Annotation of the Drosophila Melanogaster Euchromatic Genome: A Systematic Review

Research

Annotation of the *Drosophila melanogaster* euchromatic genome: a systematic review

Sima Misra*†, Madeline A Crosby‡, Christopher J Mungall†§, Beverley B Matthews‡, Kathryn S Campbell‡, Pavel Hradecky‡, Yanmei Huang‡, Joshua S Kaminker*†, Gillian H Millburn¶, Simon E Prochnik*†, Christopher D Smith*†, Jonathan L Tupy*†, Eleanor J Whitfield¥, Leyla Bayraktaroglu‡, Benjamin P Berman*, Brian R Bettencourt‡, Susan E Celniker#, Aubrey DNJ de Grey¶, Rachel A Drysdale¶, Nomi L Harris†#, John Richter§, Susan Russo‡, Andrew J Schroeder‡, ShengQiang Shu*†, Mark Stapleton#, Chihiro Yamada¶, Michael Ashburner¶, William M Gelbart‡, Gerald M Rubin*†§# and Suzanna E Lewis*†

Addresses: *Department of Molecular and Cell Biology, University of California, Life Sciences Addition, and †FlyBase-Berkeley, University of California, Berkeley, CA 94720-3200, USA. ‡FlyBase-Harvard, Department of Molecular and Cell Biology, Harvard University, Biological Laboratories, 16 Divinity Avenue, Cambridge, MA 02138-2020, USA. §Howard Hughes Medical Institute, University of California, Berkeley, CA 94720, USA. ¶FlyBase-Cambridge, Department of Genetics, University of Cambridge, Downing Street, Cambridge CB2 3EH, UK. ¥EMBL Outstation, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. #Department of Genome Sciences, Lawrence Berkeley National Laboratory, One Cyclotron Road Mailstop 64-121, Berkeley, CA 94720, USA.

Correspondence: Sima Misra. E-mail: sima@fruitfly.org.

Published: 31 December 2002

Genome ***Biology*** 2002, **3(12)**:research0083.1–0083.22

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2002/3/12/research/0083

Received: 16 October 2002
Revised: 28 November 2002
Accepted: 28 November 2002

(Print ISSN 1465-6906; Online ISSN 1465-6914)

Abstract

Background: The recent completion of the *Drosophila melanogaster* genomic sequence to high quality and the availability of a greatly expanded set of *Drosophila* cDNA sequences, aligning to 78% of the predicted euchromatic genes, afforded FlyBase the opportunity to significantly improve genomic annotations. We made the annotation process more rigorous by inspecting each gene visually, utilizing a comprehensive set of curation rules, requiring traceable evidence for each gene model, and comparing each predicted peptide to SWISS-PROT and TrEMBL sequences.

Results: Although the number of predicted protein-coding genes in *Drosophila* remains essentially unchanged, the revised annotation significantly improves gene models, resulting in structural changes to 85% of the transcripts and 45% of the predicted proteins. We annotated transposable elements and non-protein-coding RNAs as new features, and extended the annotation of untranslated (UTR) sequences and alternative transcripts to include more than 70% and 20% of genes, respectively. Finally, cDNA sequence provided evidence for dicistronic transcripts, neighboring genes with overlapping UTRs on the same DNA sequence strand, alternatively spliced genes that encode distinct, non-overlapping peptides, and numerous nested genes.

Conclusions: Identification of so many unusual gene models not only suggests that some mechanisms for gene regulation are more prevalent than previously believed, but also underscores the complex challenges of eukaryotic gene prediction. At present, experimental data and human curation remain essential to generate high-quality genome annotations.

Background

In the lexicon of genomics, an annotation is any feature tied to the genomic DNA sequence, for example, a protein-coding gene model, a transposon, or a non-protein-coding RNA gene. Adding such annotations to the sequence of a genome in a rigorous and consistent way is a prerequisite for the efficient use of that sequence in biological research. Learning how to identify, display, query, and interpret genome features in well-characterized model organisms like the fruit fly, *Drosophila melanogaster*, is crucial to understanding the genomes of more complex organisms, including *Homo sapiens*.

A major long-term goal of the FlyBase [1,2] annotation project is to overlay the *Drosophila melanogaster* genomic sequence with all available biological information and to provide traceable evidence for every annotation in a publicly accessible database. In this paper, we provide a description of our most recent step toward these goals.

In March 2000, a collaborative group including Celera Genomics, the Berkeley and European *Drosophila* Genome Projects (BDGP and EDGP), and a number of additional *Drosophila* experts published the annotated, nearly finished genomic sequence of the fruit fly [3,4]. This annotated sequence was called Release 1, in anticipation of future changes to the sequence and annotations. At that time, the annotation of genes relied heavily on computational gene-prediction algorithms with only limited human curation. The BDGP provided approximately 80,000 expressed sequence tags (ESTs), mostly from the 5′ ends of genes, which were used in the computational analyses of the genome [5]. Because these ESTs were derived from non-normalized cDNA libraries and were limited in number, they corresponded to only about 40% of all genes in the genome [5]. Complete or nearly complete sequences for an overlapping set of approximately 2,500 known *Drosophila* genes in GenBank/EMBL/DDBJ were also available [3]. Owing to the nature of whole-genome shotgun (WGS) assembly, the 1,630 gaps present in the genome tended to occur at the sites of repetitive sequence [3]; gaps corresponding to transposable elements were filled with composite sequences (reflecting sequence reads from throughout the genome) rather than the actual sequence. Release 1 predicted 13,601 protein-coding genes, encoding 14,080 transcripts; each gene was assigned a unique CG identifier. The coordinates and predicted sequences of the annotations, although not the evidence for the predictions, were made available to GenBank/EMBL/DDBJ [6-11] and FlyBase, the public databases charged with making these annotations accessible to the research community. In FlyBase, the annotations were made available as part of the genome annotation database, Gadfly [12].

Release 2, a collaborative effort between Celera Genomics and the BDGP, was submitted to GenBank/EMBL/DDBJ and FlyBase in October 2000, after approximately 330 of the gaps in the Release 1 sequence had been filled. Changes to the annotations were based largely on approximately 6,000 new 3′ ESTs sequenced by the BDGP, which increased the number of genes with 3′ UTRs and allowed further refinement in gene structures. Sequences of transposable elements remained inaccurate, being based on composite sequences. In all, 748 transcripts were modified, 114 transcripts were deleted, and 336 transcripts were added. Release 2 predicted 13,474 protein-coding genes, encoding 14,335 polypeptides, of which 13,218 (92%) were unchanged relative to Release 1. Thus, the change from Release 1 to Release 2 was minimal.

Inaccuracies in the Release 1 and 2 predicted gene structures resulted mainly from computationally predicted annotations which lacked supporting cDNA data. In addition, the annotation was carried out rapidly by a large and diverse group of curators. Mistakes in the annotation of more than 1,000 genes were reported to FlyBase in error reports from the community, and over 1,000 discrepancies between the translated annotations and those in the curated protein database SWISS-PROT [13] were reported by Karlin *et al.* [14]. Finally, a report of 1,042 new predicted annotations that did not match any of the original 13,601 predicted genes [15], and another based on analysis of testes cDNA sequences [16], suggested that the initial annotation may have missed a substantial number of genes.

The *D. melanogaster* 116.8 megabase (Mb) euchromatic genomic sequence has now been finished to high quality [17]. Here we report the results of the re-evaluation of previous annotations in light of the finished euchromatic genome and considerable additional experimental data. We call this sequence and new annotation set Release 3.

To support this re-annotation effort, a computational 'pipeline' was created, and the results were stored in a new Gadfly database, so that evidence for the annotations can be tracked and queried by the public [12]. To identify new features in the genome, we utilized prediction software and annotated alignments of non-protein-coding genes, transposons [18], and pseudogenes. To improve the extent and consistency of human curation, a small group of expert FlyBase curators visually inspected each gene in the entire euchromatic sequence, using defined rules to integrate computational analyses, cDNA data and protein alignments into updated annotations. To assess the accuracy of the exon-intron structures, we compared the resulting annotations to the subset of curated peptides in SWISS-PROT and TrEMBL that are based on experimental evidence [12].

The annotations in Release 3 alter the majority (85%) of gene models, yet confirm that previous releases accurately reflected the number of protein-coding genes. The gene models have been enhanced in a number of ways. The number of genes with annotated untranslated regions (UTRs) and alternative transcripts has increased as a direct

result of the increase in EST and complete cDNA sequences, and the fine details of the exon-intron structure are significantly improved. Numerous genes have been merged and/or split - that is, the partitioning of adjacent exons into individual gene models has changed - based on cDNA and protein sequence alignments. Overall, the improved annotations result in changes in more than 40% of the predicted proteins; however, more than 85% of the exons in the originally predicted genes contain sequences that are present in predicted exons in Release 3. We describe these changes under the headings 'Genome statistics: how is Release 3 different?', 'New and deleted annotations', and 'Structural changes to gene models' in Results and discussion.

The new annotations reveal a surprising number of genes that fall outside the typical definition of a protein-coding gene model with a 5′ UTR, coding sequence (CDS), and 3′ UTR distinct from neighboring genes. We found genes containing 3′ UTR sequences that overlap the 5′ UTR of the gene immediately downstream, examples of dicistronic transcripts (two or more distinct and non-overlapping coding regions contained on a single processed mRNA), and genes that, by means of alternative splicing, encode two completely distinct non-overlapping peptides. These atypical gene models illustrate the complexity of detailed annotation and pose new challenges for the computational annotation of genomic sequence. We describe these unusual genes, as well as assessment of and access to the data, under the headings 'Complex gene models', 'Assessment of Release 3 quality', 'Accessing data and reporting errors', and 'Future updates'.

Results and discussion

We developed a set of rules for annotation to help curators using the Apollo annotation tool [19] to move quickly through the computational results for each gene, and to annotate gene models as consistently as possible (see Materials and methods). Curators predicted transcripts supported by some combination of: computational gene structure predictions made by the Genie [20] and GENSCAN [21] programs; sequence similarities to proteins in flies and other species detected with BLASTX protein similarities, or TBLASTX similarities to virtually translated cDNA sequences [22,23]; and alignments of *Drosophila* ESTs and full-insert cDNA sequences generated by Sim4 [24] (see Materials and methods). Computational results overlapping transposon annotations were ignored when annotating protein-coding genes and RNAs; transposable elements were annotated separately [18].

We report here the re-annotation and analysis of the euchromatic portion of the *D. melanogaster* genome. There is no universally accepted definition of heterochromatin versus euchromatin; hence any declared boundary is somewhat arbitrary. We have adopted the following operational distinction: the 116.8 Mb sequence in the Release 3 large chromosome arm contigs constitutes euchromatin and is the subject of this report. The 20.7 Mb of sequence in the whole-genome shotgun-3 (WGS3) assembly [17] that is not represented in the large chromosome-arm contigs constitutes heterochromatin; analysis of these sequences is reported in an accompanying paper [25]. However, we note that this is an oversimplification, as the proximal portions of the large chromosome arm sequences extend into what is defined as heterochromatin by cytological criteria (see [25] for a detailed description). The chromosome arm contigs are essentially finished, high-quality sequences, whereas the WGS3 non-redundant contigs are draft quality [17]. The euchromatic regions contain 98% of known genes and the statistics provided in Tables 1-4 refer only to these genes. The 2% of genes found in heterochromatin cannot be annotated with sufficient confidence to provide this detailed information, because the WGS3 is still draft sequence. However, the addition of these genes is unlikely to appreciably change the results of our analysis.

Genome statistics: how is Release 3 different?

Increase in the number of exons and transcripts, but not genes

Although the re-annotation process changed the majority of gene models, the number of protein-coding genes changed minimally, from 13,601 genes in Release 1 to 13,474 genes in Release 2 to 13,676 in Release 3, of which 13,379 are in the euchromatin (Table 1) and 297 in the heterochromatin [25]. However, the Release 3 gene structures have changed to contain more exons. The total number of unique exons in euchromatin, defined as having unique sequence coordinate termini, has increased 11% from 54,793 in Release 2 to 60,897 in Release 3 (see Table 1). The number of protein-coding exons has increased as well, from 50,667 to 54,934 (we define a protein-coding exon here as any exon containing CDS, even if the majority of the exon is UTR). The consequence is that the average number of exons per gene has increased from 4.1 in Release 2 to 4.6 in Release 3, which is very similar to *C. elegans* (4.5 [26]) and *Arabidopsis* (4.6 [27]) but significantly lower than *H. sapiens* (8.9, see, for example [28]).

A major contributor to the increase in exons is the increase in the number of protein-coding genes with identified 5′ UTRs. One limitation of *ab initio* gene prediction programs is that they predict only open reading frames (ORFs): EST and full-length cDNA data are absolutely essential to identify UTRs. The expanded set of available EST/cDNA data led to a significant increase in the number of annotated genes and transcripts with 5′ UTRs (from 57% of the genes in Release 2 to 76% in Release 3) and 3′ UTRs (from 36% of the genes in Release 2 to 72% in Release 3; Table 1). These numbers reflect data availability: sequences from cDNA clones representing at least one transcript from approximately 78% of the genes in *Drosophila* were supplemented by a large number of additional 5′ ESTs [29,30]. The length of the UTRs also increased with these new data (Table 1):

Table 1

Comparison of Release 2 and 3 genome statistics

Description*	Release 2 (% of total)	Release 3 euchromatin† (% of total)
Total protein-coding genes	13,474	13,379
Total length of euchromatin	116.2 Mb	116.8 Mb
Exons	54,793	60,897
Protein-coding exons‡	50,667	54,934
Length of genome in exons	23.3 Mb (20%)	27.8 Mb (24%)
Introns	41,381	48,257
Genes with 5′ UTR	7,680 (57%)	10,227 (76%)
Transcripts with 5′UTR	8,499 (59%)	14,707 (81%)
Average 5′ UTR length	204 nucleotides	265 nucleotides
Genes with 3′ UTR	4,824 (36%)	9,646 (72%)
Transcripts with 3′ UTR	5,381 (38%)	14,012 (77%)
Average 3′ UTR length	370 nucleotides	442 nucleotides
Average ratio of length of CDS/transcript§	0.86	0.75
Total protein-coding transcripts	14,335	18,106
Genes with alternative transcripts	689 (5%)	2,729 (20%)
Average number of transcripts per alternatively spliced gene	2.25	2.75
Total number alternative transcripts	861	4,743
Number of introns contained in 5′UTRs	2,977	6,787
Number of introns contained in 3′ UTRs	1,004	1,088
Unique peptides¶	13,922	15,848
Unique peptides unchanged from R2 to R3	8,769 (63%)	8,769 (55%)
Genes deleted from R2 to R3	345	NA
New protein-coding genes in R3	NA	802

*Abbreviations: UTR, untranslated region; CDS, (protein)-coding sequence; R2, Release 2; R3, Release 3; NA, not applicable. All statistics are for protein-coding genes only. †Based on the annotation of protein-coding genes in the euchromatin (long chromosome arms); another 297 protein-coding genes are annotated in the heterochromatin (non-redundant WGS3 [25]). In this and Tables 2-4, the numbers are based on a version of the annotation database frozen on November 25, 2002. ‡Any exon containing CDS, even if the majority of the exon is UTR. §The length of the coding region divided by the length of the entire protein-coding transcript, averaged over all protein-coding transcripts. ¶Determined because many alternative transcripts encoded the identical CDS and differed only in the UTR.

Table 2

Types of annotations in Release 3 euchromatin

Description	Release 2	Release 3
Protein-coding genes	13,474	13,379
tRNA genes	0	290
microRNA genes	0	23
snRNA genes	0	28
snoRNA genes	0	28
Pseudogenes	0	17
Miscellaneous non-coding RNA	0	38
Transposons	0	1,572
Total annotations	13,474	15,375

the average 5′ UTR length per transcript (for genes with annotated UTRs) increased by 30%, to 265 nucleotides, and the average 3′ UTR length (for genes with UTRs) by 19%, to 442 nucleotides.

Four times as many genes in Release 3 (20%) as compared to Release 2 (5%) show alternative transcripts (Table 1). The vast majority of these are due to alternative splicing (an introduced bias; see Materials and methods), but 13% are due to alternative promoters and 6% to alternative polyadenylation sites. Alternative splicing results in the 26% increase in the number of protein-coding transcripts, and is largely responsible for a 14% increase in the number of unique protein species: from 13,922 in Release 2 to 15,848 in Release 3.

Forty-five percent of predicted proteins differ from Release 2

The changes in gene models also result in larger proteins. Proteins in Release 3 have a mean length of 552 amino acids and a median of 421 amino acids. This is an increase compared to Release 2, where the mean was 503 amino acids and the median 385 amino acids. The longest transcript and protein are encoded by *dumpy* (*dp*), which encodes a massive 69.7 kilobase (kb) mRNA and a 23,054 amino-acid polypeptide. The Dp protein is a component of the extracellular matrix, and appears to serve as an elastic adhesion molecule at cuticle-cell junctions, such as the epidermal-cuticle interface [31].

The vast majority (94%) of the Release 3 annotations contain sequences that are present in exons from Release 2; however, only 63% of the unique peptides in Release 2 are unchanged in Release 3 (Table 1). Of the 15,848 unique Release 3 peptides, 8,769 (55%) are exact matches to Release 2 peptides. Reciprocally, 45% of the peptides are different from Release 2, emphasizing that, although the overall picture of the number and distribution of transcription units in the *D. melanogaster* genome remains largely the same, the new annotations include many changes to the protein products encoded by the genome.

New and deleted annotations

The re-annotated genome now includes non-protein-coding genes (tRNAs, microRNAs, snRNAs, and snoRNAs) and transposable elements. Although some of these features were reported in the publication of the first release [3], the coordinates of these features were not included in data sent to public databases.

Table 3

Evidence supporting the euchromatic protein-coding gene models*

Data category	Number of Release 3 protein-coding genes	% of Release 3 protein-coding genes
Total	13,379	100
Release 2 annotations	12,549	94
Gene-prediction data only	815	6
Genie gene predictions	12,427	93
GENSCAN gene predictions	12,853	96
BLASTX/TBLASTX homologies	10,996	82
ESTs and DGC cDNA sequencing reads	10,406	78
GenBank accessions†	3,104	23
ARGS (RefSeq)‡	795	6
Error report submissions	825	6
Full-insert DGC cDNAs§	9,297	69

*Determined by assessment of alignment overlap of data category versus gene model. †Not including those contributed by the BDGP (not mutually exclusive categories: many of these genes also have representative cDNA clones in the DGC). ‡ARGS, annotated reference gene sequence; high-quality FlyBase gene-level annotations that include data from the published literature; contributed to the NCBI reference sequence (RefSeq) project. §For a rigorous assessment of the quality of the DGC cDNAs, see [30].

Table 4

Classification of euchromatic transcript and gene confidence values

Confidence value*	Number of transcripts (%)	Number of genes† (%)
1	1,227 (7%)	1,201 (9%)
2	2,098 (12%)	1,975 (15%)
3	3,122 (17%)	2,437 (18%)
4	11,659 (64%)	7,766 (58%)
Total	18,106	13,379

*Confidence values reflect number of types of supporting data, from 1 (lowest) to 4 (highest); see Materials and methods. †Genes were assigned the confidence value of the highest-scoring transcript.

Transposable elements

The sequences of the vast majority of transposons in Releases 1 and 2 were composites derived from a number of copies of that transposon type. In Release 3, these composite sequences are replaced with the actual sequences present in the sequenced *y*^{1}*; cn*^{1} *bw*^{1} *sp*^{1} strain for each individual element [17,18]. In all, 1,572 transposons are annotated in the euchromatic Release 3 genome (Table 2): 682 long terminal repeat (LTR) transposons, 486 LINE transposons, 372 terminal inverted repeat (TIR) transposons, and 32 foldback (FB) elements. These data include both full-length and partial elements. Details of these analyses are reported in an accompanying article [18].

Non-protein-coding RNA genes

Small, non-protein-coding RNAs are also included in this re-annotation. We searched for new tRNA genes using the program tRNAscan-SE [32] and Sim4 alignments to known tRNAs: 290 are annotated in the euchromatin (Table 2). Release 1 reported 292 tRNAs [3]; two tRNA genes were deleted as a result of sequence finishing resolving repeated regions of the genome.

Other non-protein-coding RNAs are limited, in general, to those already curated in the FlyBase database [1,2]. All 23 of the known microRNAs in *Drosophila* are located precisely in the Release 3 genome. We annotated the majority of the 45 small nuclear RNAs (snRNAs) involved in splicing, with the exception of the four snRNAs, K2a, K2b, K8, and K9 [33], for which there were no sequence or cytological data available. Of the 41 snRNAs supported by such data, we found that nine were redundant entries, and another five could not be identified at the previously specified cytological locations, possibly due to strain variation and/or inaccuracy in previous localization experiments. Thus, we precisely located by sequence alignment 28 snRNAs in the genome, including a new copy of the *snRNA:U4atac* gene in the 83A region.

All nine of the small nucleolar RNA (snoRNA) genes in FlyBase were identified by Sim4 alignment of sequence obtained from the literature. In addition, we incorporated data from Tycowski and Steitz [34] and located 19 more snoRNAs. Identification of other snoRNAs should be possible in the future with the use of algorithms like Snoscan, which looks for 2′-O-ribose methylation guide snoRNAs [35]; however, the program will have to be customized for *Drosophila*.

The longer non-protein-coding RNA genes *αγ-element, bft, RNaseP:RNA, Hsr-omega, 7SLRNA, pgc, roX1, roX2,* and *iab-4*, are annotated in the genome. 27 new 'miscellaneous non-coding RNA' genes were detected by alignment of spliced DGC cDNAs that did not appear to contain an ORF of significant length. In some cases these appear to be candidate antisense genes, which have also been reported in other organisms [36]. Further experiments will be necessary to verify the existence of these interesting genes and to determine their function.

Pseudogenes

The number of pseudogenes reported in *Drosophila* is substantially smaller than that in *Caenorhabditis elegans* [37,38]. We annotated the 12 pseudogenes in FlyBase that map to the euchromatic sequence and correspond to protein-coding paralogs (see Supplementary Table 1 in the additional data files). We identified five new pseudogenes: four histones and one lectin (*CR31541*) (Supplementary

Table 1). Of these 17 pseudogenes, 15 are recombinationally derived (with introns, in tandem to their functional paralogs), one (*Mgstl-Psi*) is retrotransposed (with a poly(A) tail, lacking introns which its functional paralog possesses), and one is too degenerate to classify definitively. We did not make any attempt to comprehensively survey for new retrotransposed pseudogenes or annotate pseudogenes identified by Echols *et al.* [37]. WormBase [39,40] currently reports 392 pseudogenes in *C. elegans*. It is very likely that a subset of the genes identified as protein-coding genes in Release 3 are actually pseudogenes. In particular, 19 protein-coding genes were noted as containing a 'probable mutation in the sequenced strain' and more than 400 were marked 'problematic' because of inconsistencies with the experimental evidence and the predicted ORFs.

New protein-coding genes

Release 3 contains a total of 802 new protein-coding genes (Table 1), that is, gene models that show no overlap with exons in Release 2. Of these, 55 (7%) are based solely on gene-prediction data, and 20 of these 55 are based on GENSCAN predictions alone. Unlike Releases 1 and 2, which relied heavily on Genie [3], Release 3 annotations did utilize GENSCAN predictions (with at least one exon with a score > 45) in the absence of other data. The majority of the new genes show matches to EST (573; 71%) or full-insert cDNA sequences (273; 34%), indicating the importance of these alignments in identifying new genes missed by the *ab initio* gene prediction programs. An additional set of new genes was identified by the community in error reports (52; 7%) or in GenBank/EMBL/DDBJ submissions (58; 7%). Finally, we created 338 (42%) new annotations in Release 3 using protein homology data from BLASTX analysis, arising from the comparison of translated Release 3 sequence with sequence of other proteins in *Drosophila* or other model organisms, in the absence of other supporting data.

Release 3 sequence 'finishing' had the largest impact on areas of repetitive sequence, because the Release 2 WGS sequence assembly often collapsed these regions [17]. Duplicated sequences present assembly challenges to genome sequencing efforts; tandemly duplicated genes tend to collapse in sequence assembly and cannot be annotated until the duplications are resolved. Whole-genome analysis of the Release 2 sequence suggested that *Drosophila* has fewer newly duplicated genes than nematodes or yeast [41]. We investigated whether sequence finishing might have uncovered previously undiscovered duplicated genes in *Drosophila*. From this analysis, we found that the number of newly duplicated genes is more similar to *C. elegans* and *Saccharomyces cerevisiae* than previously believed.

We measured the frequency of newly annotated duplicated genes by comparing each of the transcripts encoded by the 802 new genes in Release 3 to all Release 3 transcripts using the BLASTN program. Of the new genes, 124 (15%) have duplicate genes (75% identity, probability = 1 x e^{-25}) somewhere in the genome (whereas 10% of a random sample of *Drosophila* genes have duplicates by this measure). Thirty-six new genes are in repeat regions that were collapsed in Release 2 and have now been resolved. For example, in the previously annotated ten-gene trypsin cluster on chromosome arm 2R, three new trypsin genes (*CG30025*, *CG30028*, *CG30031*) have been added ([17] and Figure 1).

Deleted protein-coding genes

We rejected a total of 345 Release 2 genes during the Release 3 re-annotation (Table 1), primarily on the basis of a lack of supporting computational or experimental evidence (see Materials and methods). Nineteen Release 2 genes were deleted because they were contained within transposable elements. If based solely on computational gene-prediction evidence, genes that were less than an arbitrary length of 100 amino acids were deleted. We required an arbitrary length of 50 amino acids for all annotations not specifically supported by literature references (for example, the DIRG genes [42]), and 42 of the deleted Release 2 annotations were removed because they failed to meet this criterion (Figure 2a, inset).

The sizes of the predicted protein products in Release 2 and Release 3 were compared (Figure 2a), along with the protein sizes of Release 2 annotations deleted in Release 3, and sizes of proteins newly added in Release 3 (Figure 2b). When examining the size range of 0-50 amino acids, there is a marked decrease in Release 3 annotations compared to Release 2, due to the more stringent data requirements for small annotations in Release 3 (Figure 2b, inset).

Structural changes to gene models

There were several major categories of changes to gene models: adjustment of exon boundaries, especially at the 5′ and 3′ ends of genes; deletion or addition of exons; merges of two or more genes; splitting of genes; and gene splits/merges, in which neighboring or nested gene models were split and the exons from the original gene models were redistributed between the updated models.

The majority of changed gene models fall into the first two categories: adjusted exon boundaries or deleted or added exons. Many of these changes affect only UTR sequences, leaving the CDS unchanged. A small but significant number of gene models were more complicated and involved exon redistribution. When these genes were merged and/or split, new CG identifiers were assigned to indicate a substantial change to the gene models.

Gene merges

Evidence supporting the merger of gene models came mainly from the alignment of full-length cDNA sequences and, to a lesser extent, from protein homology evidence.

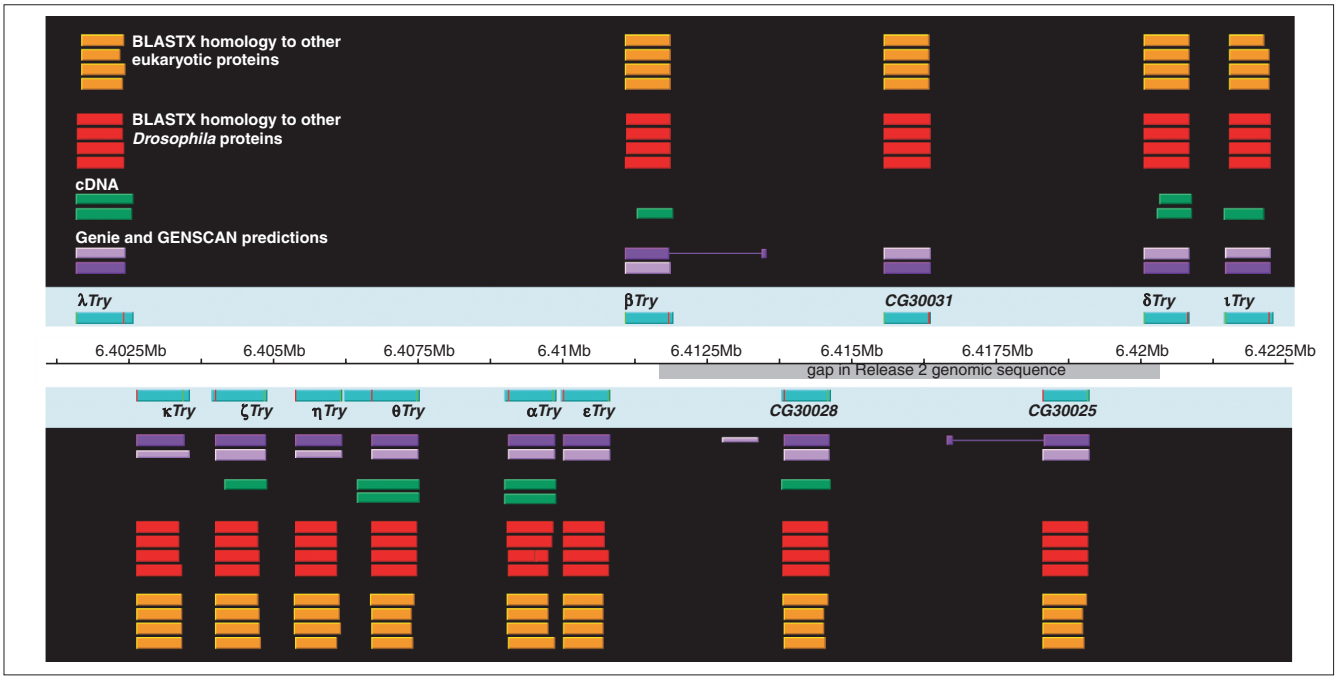

Figure 1
A resolved misassembly from Release 2 sequence contains new trypsin genes. This illustration and Figures 3-8 are derived from the output of the graphical annotation tool Apollo [19], but these illustrations are not intended to be a direct representation of the data used to annotate the regions. Only evidence (shown in the black panels) directly used to annotate the gene models (shown in the cyan panels) are depicted in these illustrations. The plus strand is shown above the center scale, the minus strand below the center scale. Thin lines represent introns and thick boxes represent exons. Vertical green lines in the exons represent start codons and vertical red lines represent stop codons. An 8.5-kb region of genomic sequence on chromosome arm 2R was missing in Release 2 because of an apparent misassembly that incorrectly joined two tandemly repeated trypsin genes with a concomitant deletion of the intervening sequence (region shown in gray in the center scale). The missing sequence constituted an inverted repeat of 4kb bordered by a simple repetitive sequence (S.C., unpublished results). Resolution of this error in Release 3 has led to the annotation of three new trypsin genes (blue rectangles): *CG30025* (similar to *βTry*), *CG30028* (similar to *γδTry*), and *CG30031* (similar to *γδTry*). Gene-prediction data (dark purple for Genie and lavender for GENSCAN), cDNA data (dark green), and BLASTX protein similarity (red for *Drosophila* proteins, orange for other species' proteins) support the new trypsin genes.

Merges based solely on BLASTX similarity were more difficult, as the exact exon-intron structure of the merged model was not experimentally indicated. A total of 1,351 Release 2 genes were merged to form 602 (5% of total) Release 3 genes. Sometimes the original predictions were spaced quite far apart in the genome, a probable reason that the gene prediction algorithm(s) separated the exons. For example, multiple ESTs support a merge of *CG14409* and the *Flotillin-2* gene (*Flo-2* or *CG11547*), adding two 5′ exons almost 80 kb away from the Release 2 annotation of the *Flo-2* gene (Figure 3). The new *Flo-2* transcript encodes a protein with an additional 50 amino acids at its amino terminus.

Gene splits

Gene model splits were often necessitated by the facts that gene-prediction programs such as Genie and GENSCAN can string together genes that lie close to each other and do not resolve nested genes. Of the Release 2 genes, 322 were split to form 675 (5% of total) Release 3 genes. For example, the annotated gene *CG6645*, with 5 exons in Release 2 (Figure 4), appears to have been based on a Genie prediction (note that GENSCAN had predicted two separate genes). EST evidence and BLASTX homology to other fly proteins indicated that this gene should be split into two three-exon genes, *CG32054* and *CG32053*. One 5′ UTR exon and one protein-coding exon in *CG32053* were missed by both Genie and GENSCAN. Thus, neither Genie nor GENSCAN correctly predicted the structure of these two genes; each correctly predicted aspects of the gene models, but EST and BLASTX data were necessary to accurately determine the structure of the two genes.

Gene splits/merges

Gene splits/merges were defined as changes involving more than one gene in both Release 2 and 3. While not common, splits/merges are interesting in that they involve simultaneous restructuring of multiple Release 2 annotations. One notable example is the split of *CG8278* into the *CG30350* and *sns* annotations (Figure 5). In this instance, BLASTX, GenBank/EMBL/DDBJ, and cDNA records indicate that the 3′ half of *CG8278* should be split off as a separate gene model (*CG30350*), while the GenBank:AF254867

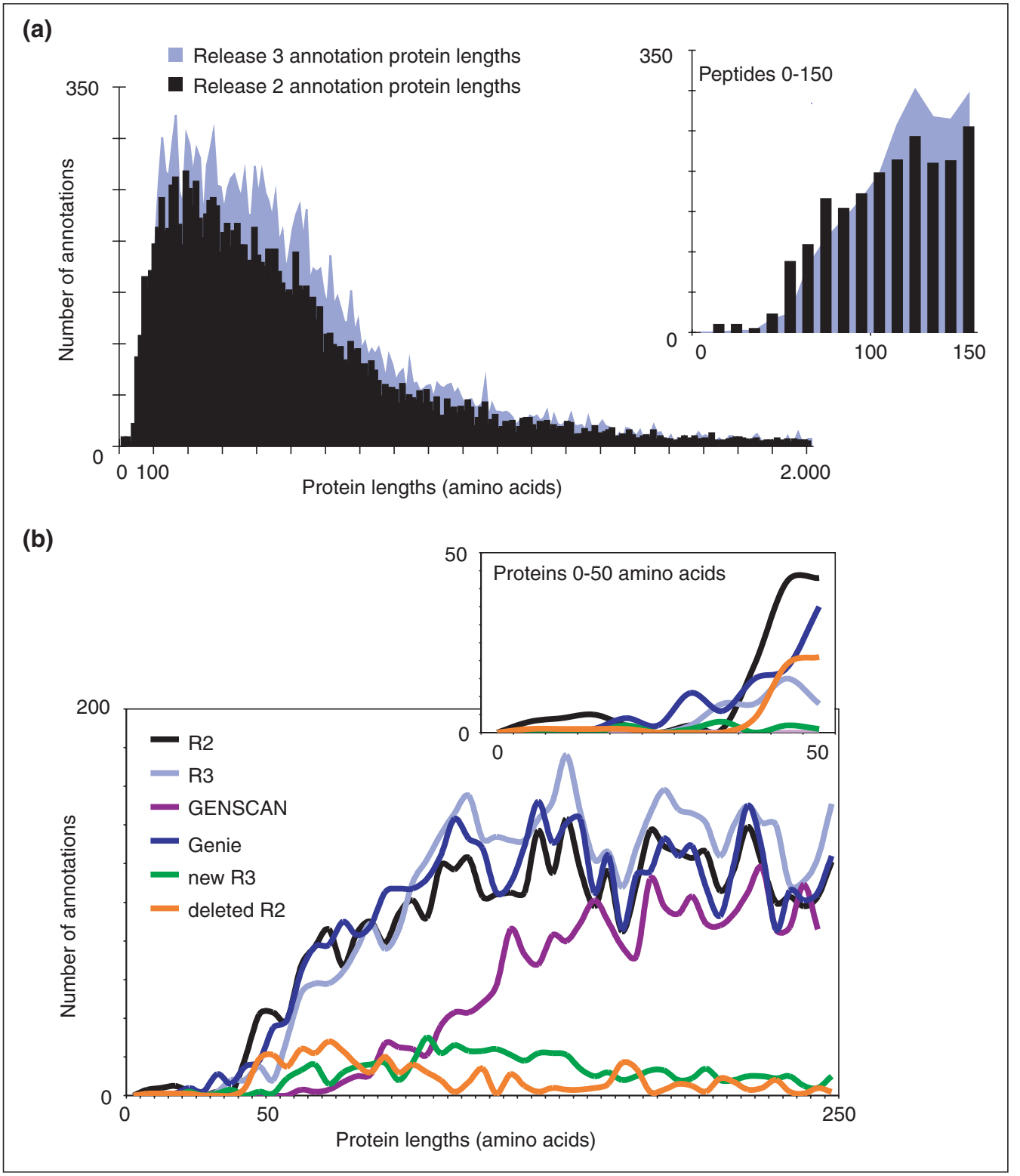

Figure 2
Distribution of predicted peptide lengths in Release 2 and 3. **(a)** Comparison of protein lengths less than 2,000 amino acids shows that overall, Release 3 proteins of all lengths (blue) are more numerous than those in Release 2 (black). One exception is those proteins shorter than 100 amino acids: because of stricter data requirements for Release 3 annotations, some small Release 2 annotations were not preserved (inset). **(b)** Comparison of Release 2 (black) and 3 (light blue) protein lengths with predictions by GENSCAN (purple) and Genie (dark blue). Also shown are the lengths of proteins that were deleted (orange) or added (green) in Release 3. Of note is the underprediction of genes expressing small proteins by the program GENSCAN (purple).

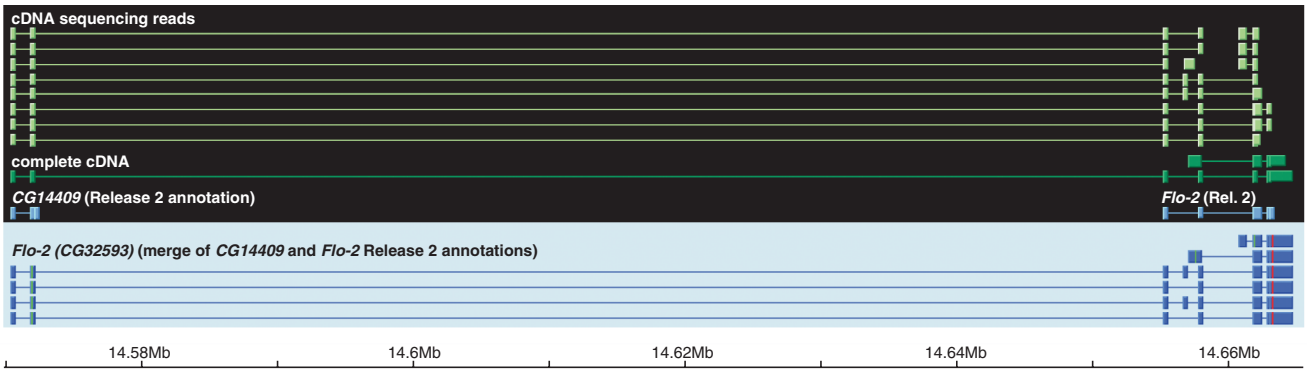

Figure 3
Release 2 annotations *CG14409* and *Flo-2* (*CG11547*) were merged to create an expanded *Flo-2* (*CG32593*) gene model. Only evidence (black panel) directly used to annotate the gene model (cyan panel) is shown. Alignments of ESTs and cDNA sequence reads (light green) and an assembled full-insert cDNA clone sequence (dark green) support the merger of the Release 2 annotation *CG14409* (light blue) and the adjacent gene, *Flo-2* (light blue), on the X chromosome. The expanded Release 3 *Flo-2* annotation (dark blue) was assigned the new annotation number *CG32593* to reflect this significant change. Predicted exons derived from a single cDNA clone are joined by thin horizontal lines, indicating introns. Predicted exons not so joined derive from different cDNA clones. Distance along the chromosome arm is shown in the scale at the bottom; the scale is black to denote the location of these annotations on the plus strand. Although the lowermost two transcripts appear to be duplications of other transcripts, they contain a slight variation in their 5′ exon that is not visible at the scale used in this figure.

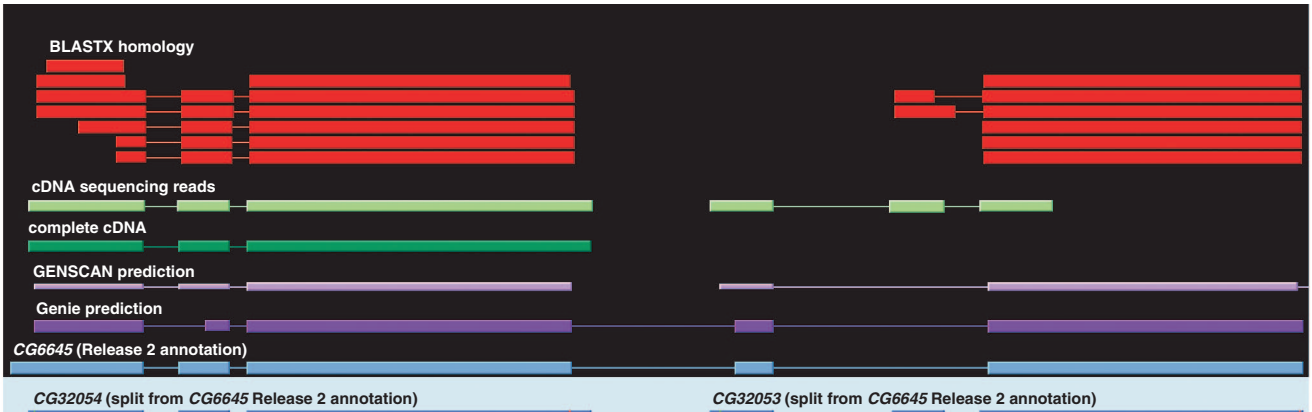

Figure 4
The Release 2 annotation *CG6645* was split to create *CG32054* and *CG32053*. Only evidence (black panel) directly used to annotate the gene models (cyan panel) is shown. While Release 2 annotation *CG6645* on chromosome arm 2L consisted of a single long transcript (light blue), review of assembled EST and cDNA sequencing reads (light green) and BLASTX evidence (red) led to the creation of two smaller Release 3 annotations from the two halves of the original gene model. These new annotations (dark blue) were designated *CG32054* and *CG32053*. Although the Genie prediction (purple data on black panel) supports a single coding transcript, the remaining data were judged to be stronger evidence of two separate genes. Note that for *CG32053*, the second exon was not included in either gene prediction, and was added on the basis of on cDNA sequencing read and BLASTX evidence (arrow). The chromosome scale at the bottom is red to denote the location of these annotations on the minus strand.

record indicates that the 5′ exon of *CG8278* plus six other Release 2 annotations should be merged into the extensive *sns* annotation. There were 93 cases in which Release 2 annotations suffered a reassignment of exons more complex than a simple gene split or merge to generate Release 3 annotations.

Complex gene models

Eukaryotic genomes defy our efforts to impose simple or computable rules of gene structure and organization. We discovered many examples of genes that overlap, that share transcription units, or that produce a dozen or more different protein products. FlyBase uses the following nomenclature

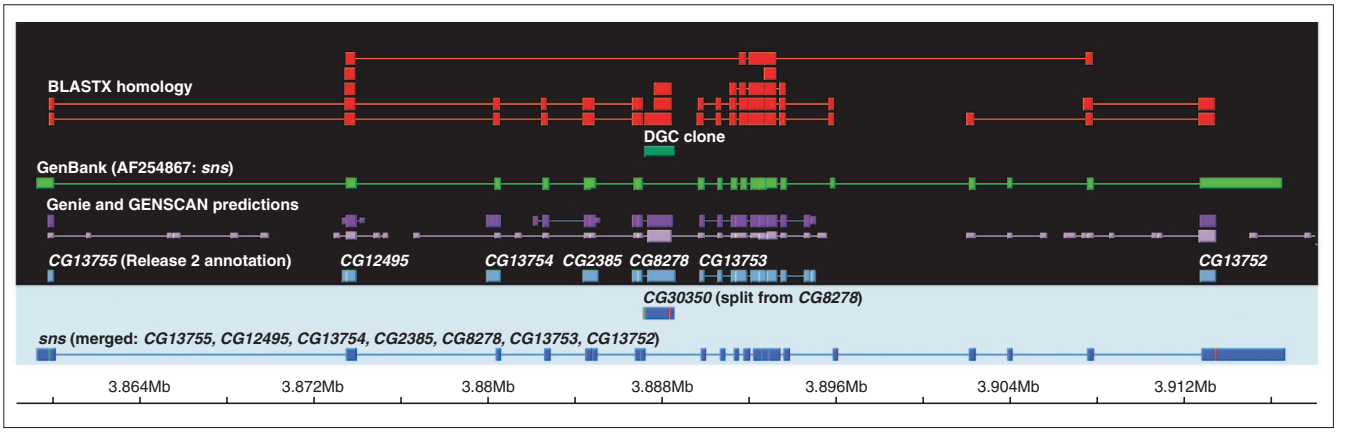

Figure 5
Complex split/merge creates updated *sns* annotation and new annotation *CG30350*. Only evidence (black panel) directly used to annotate the gene models (cyan panel) is shown. Occasionally, annotation of a particular region required complex rearrangement of the exons comprising the Release 2 gene models. In this case, the second exon of the Release 2 annotation *CG8278* (light blue) was split off as a new gene (*CG30350*, dark blue) on the strength of DGC cDNA data (dark green) and BLASTX evidence (red). The remaining exon of *CG8278*, along with six other Release 2 annotations (*CG13755*, *CG12495*, *CG13754*, *CG2385*, *CG13753*, and *CG13752*; light blue), were merged together into the large *sns* gene (dark blue), strongly supported by sequence of a full-length *sns* cDNA, GenBank:AF254867.

for complex genes: in cases of more than one transcript derived from the same genomic region (and from the same DNA strand), FlyBase assigns gene designations based on the extent of the coding regions, not the extent of the transcripts. If there is any overlap within the protein products produced, even (theoretically) a single amino acid, FlyBase considers those proteins to be products of a single gene. Alternative splicing or dicistronic transcripts may result in completely non-overlapping protein products produced from overlapping transcripts; these are described in FlyBase as separate genes. An interesting example is the previously described *Su(var)3-9* gene [43], which encodes different transcripts that share 5′ coding exons; these overlapping transcripts encode two functionally different proteins, one a chromatin-binding factor and the other a translation-initiation factor. Despite their disparity in function, the two proteins share 80 amino acids at their amino termini and are thus classified as a single gene by FlyBase. In the following sections, we describe the complex gene models we observed: nested genes, overlapping genes, alternatively transcribed genes, and dicistronic genes.

Nested genes

The phenomenon of genes within genes, in which a gene is included within the intron of another gene, is common. In the analysis of the 2.9 Mb *Adh* region of *Drosophila*, the frequency of nested genes was reported to be approximately 7% [44]. In extending this analysis to the entire euchromatin, we find that 7.5% (1,038) of all Release 3 genes, including non-coding RNAs, are included within the introns of other genes. Of the 879 nested protein-coding genes, the majority (574) are transcribed from the opposite strand of the including gene. We observed 26 cases in which the exons and introns of a gene pair are interleaved. Transposons may also be located within the introns of genes; we observed 431 such cases.

Overlapping genes

We analyzed the mRNAs predicted for neighboring genes to find those transcripts that share common non-protein-coding genomic sequence. About 15% of annotated genes (2,054) involve the overlap of mRNAs on opposite strands. Some of these involve overlapping messages that have been previously described (for example, *Dopa decarboxylase* and *CG10561* [45]); however, the vast majority were not previously known to overlap. Complementary sequences between distinct RNAs from overlapping genes on opposite strands have previously been reported in eukaryotes and have been implicated in regulating gene expression (for reviews see [46,47]). For example, the complementary sequence shared between *Dopa decarboxylase* and *CG10561* is thought to be involved in regulating the levels of these transcripts [45]. The large number of such overlapping transcripts identified here raises the possibility that antisense interactions may not be an uncommon mechanism for regulating gene expression in *Drosophila*.

We were surprised to find over 60 cases of overlapping genes on the same strand, for which cDNA/EST data indicate that the 3′ UTR of the upstream gene overlaps the 5′ UTR of the downstream gene. In some instances, the 3′ UTR of the upstream gene extends past the postulated translation start of the downstream gene. One example of such an overlapping model is *CG9455* and *Spn1 (CG9456)*, tandem genes encoding serine protease inhibitors (Figure 6). The two gene models are individually supported by a variety of BLASTX data as well as full-insert cDNA sequences. Interestingly, the

5′ exon of the DGC cDNA clone covering the *Spn1* gene (AT24862) is entirely included in the 3′-most exon of the *CG9455* DGC cDNA clone (GH04125). The existence of overlapping genes raises many questions. Are such pairs of genes typically co-regulated? Where are the transcriptional regulatory elements for the downstream gene? What are the structural constraints on the overlapping sequences?

Alternatively transcribed genes

One mechanism for increasing potential protein and regulatory diversity is through the production of alternative transcripts. Approximately 20% of Release 3 genes have more than one predicted transcript, and this is almost certainly an underestimate. Many instances of internal alternative splicing as well as alternative polyadenylation will have been missed, as our dataset of cDNA sequences contained many more 5′ ESTs than 3′ ESTs or complete cDNAs. As cDNA collections are expanded, including those representing specific stages, tissues, and cell types, additional genes with multiple transcripts and additional protein species produced by alternative splicing will undoubtedly be identified. Despite likely underestimation, the level of alternative splicing that was observed clearly illustrates that alternative splicing is an important mechanism for generating transcript diversity in *Drosophila* (see Supplementary Table 2 in the additional data files).

Alternative splicing creates opportunities for diversity both at the level of gene regulation and of protein diversity. In Release 3, 35% of the 2,729 genes encoding multiple transcripts generate only one protein product; the transcripts differ only in their UTRs. Very commonly, these alternative transcripts vary in the location of 5′ non-coding exons, suggesting the use of alternative promoters and offering the possibility of differential regulation. The other 65% of genes with alternative transcripts encode two or more protein products, indicating that alternative splicing generates considerable protein diversity in *Drosophila*.

A large number of related proteins can be produced from a single gene by the simple substitution of a single domain. This mechanism has been taken to an extreme level in the case of *mod(mdg4),* which produces at least 29 distinct transcripts that share 5′ exons, but are alternatively spliced to an array of different 3′ exons [48,49]. Remarkably, eight of these transcripts appear to be generated by a *trans*-splicing mechanism, using variable 3′ exons encoded on the opposite strand. (Seven *trans*-spliced variants were previously reported [48,49]; our analysis suggests eight.) Although we did not find any further examples of *trans*-splicing, we did find that a similar gene, *lola,* generates at least 21 alternative transcripts (including four previously described [50]). The many *lola* transcripts also share 5′ exons, but contain one of an array of different 3′ exons. Both *lola* and *mod(mdg4)* encode families of specific RNA polymerase II transcription factors that include a BTB/POZ dimerization domain near each amino terminus [50,51]. *mod(mdg4)* has been implicated in a range of cellular and developmental processes, including chromatin insulator functions [52] and apoptosis [53], and it has been suggested that its many different isoforms underlie the pleiotropic nature of this gene [49].

Alternative splicing can produce two (or more) distinct non-overlapping protein products from a single pre-mRNA species; we identified 12 such cases (for example, *Vanaso* and *α-Spec*, see Figure 7). The mRNAs produced most commonly share 5′ UTR sequences, but may also share 3′ UTR sequences. FlyBase defines complexes of this type as two

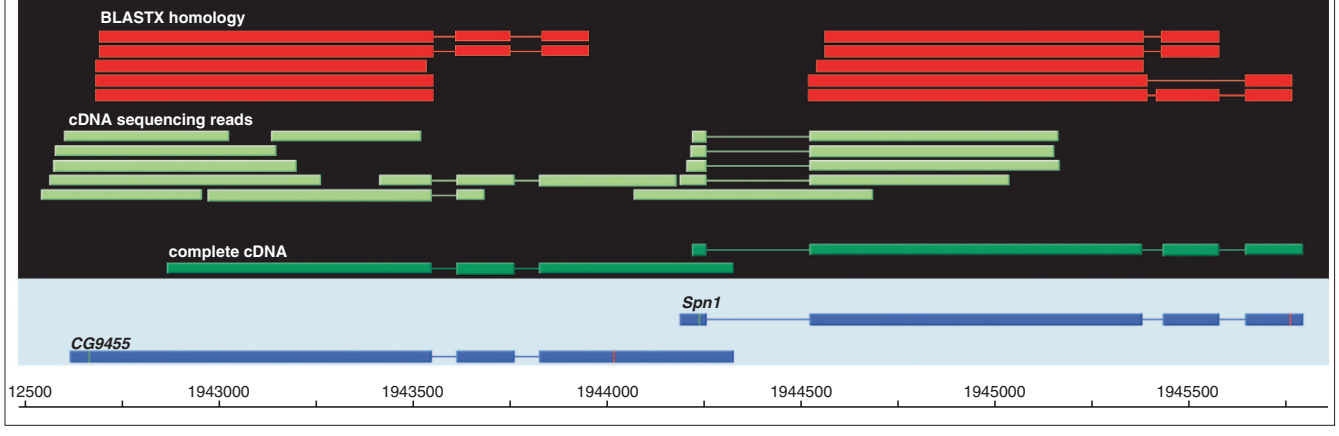

Figure 6
The 3′ UTR of *CG9455* overlaps the downstream gene *Spn1*. Only evidence (black panel) directly used to annotate the gene models (cyan panel) is shown. This example of tandem overlapping genes is supported by full-insert cDNA sequences (dark green) and assembled EST and cDNA sequencing reads (light green). The 3′ UTR of the *CG9455* transcript (dark blue) extends past the initiation site of the *Spn1* transcript (dark blue). BLASTX data (red) demonstrate that these transcripts encode independent proteins.

separate genes, since two non-overlapping protein products are produced. Although other groups sometimes describe such genes as dicistronic (since the unprocessed transcript is dicistronic), we do not include this type in our categorization of dicistronic genes (see below). The component coding regions are resolved on separate mRNAs, and thus internal translation initiation is not required. We view these cases as one extreme along a continuum of protein diversity created by alternative splicing.

Dicistronic genes

Examples of dicistronic transcripts have been previously reported in *Drosophila* [54-61]. Our results confirm that, while not common, numerous examples of apparent dicistronic transcripts are encountered in *Drosophila*. We limit the term 'dicistronic' to genes that meet the following criteria: two distinct and non-overlapping coding regions contained on a single processed mRNA, requiring internal initiation of translation of the downstream CDS. In order to categorize a transcript as dicistronic, we required that each CDS exceed 50 amino acids in length and show some similarity to known proteins. The Release 3 annotation contains 31 gene pairs that can be described as dicistronic by these criteria (Figure 8, and see Supplementary Table 3 in additional data files). This includes 12 cases for which the dicistronic transcript is represented by a single cDNA. There are 17 additional pairs, denoted as putative, for which there is insufficient BLASTX evidence to support both ORFs in a dicistronic gene model (see Supplementary Table 3). Since the determination of genes as dicistronic requires multiple classes of data to confirm the transcript structure and validate the coding regions, there are undoubtedly additional dicistronic genes yet to be uncovered throughout the genome.

For many of the predicted dicistronic genes (31/48), there is evidence supporting alternative monocistronic transcript(s) for either the upstream or downstream CDS, or for both. This includes *Mosc1A+Mosc1B*, for which the monocistronic transcript encodes a fusion protein encompassing both CDSs [59]. In some cases the dicistronic form may be less prevalent than the monocistronic forms: it has been estimated that the dicistronic *Adh+Adhr* transcript is only 5% as abundant as that of the *Adh* monocistronic transcripts [57].

Translation of the second CDS of a dicistronic transcript requires that initiation of translation occur at an internal site. There are two proposed mechanisms for the initiation of internal translation. One mechanism is that internal initiation occurs by partial disassembly of the ribosome at the termination of translation of the first CDS, followed by continued scanning by the 40S ribosomal subunit [62]. The following conditions are thought to be criteria for the ribosomal scanning mechanism: an absence of any ATG codons in the intercistronic region, an intercistronic region of 15 to 78 bp, and an optimized consensus translation start site for the second CDS. We assessed the sizes of intercistronic

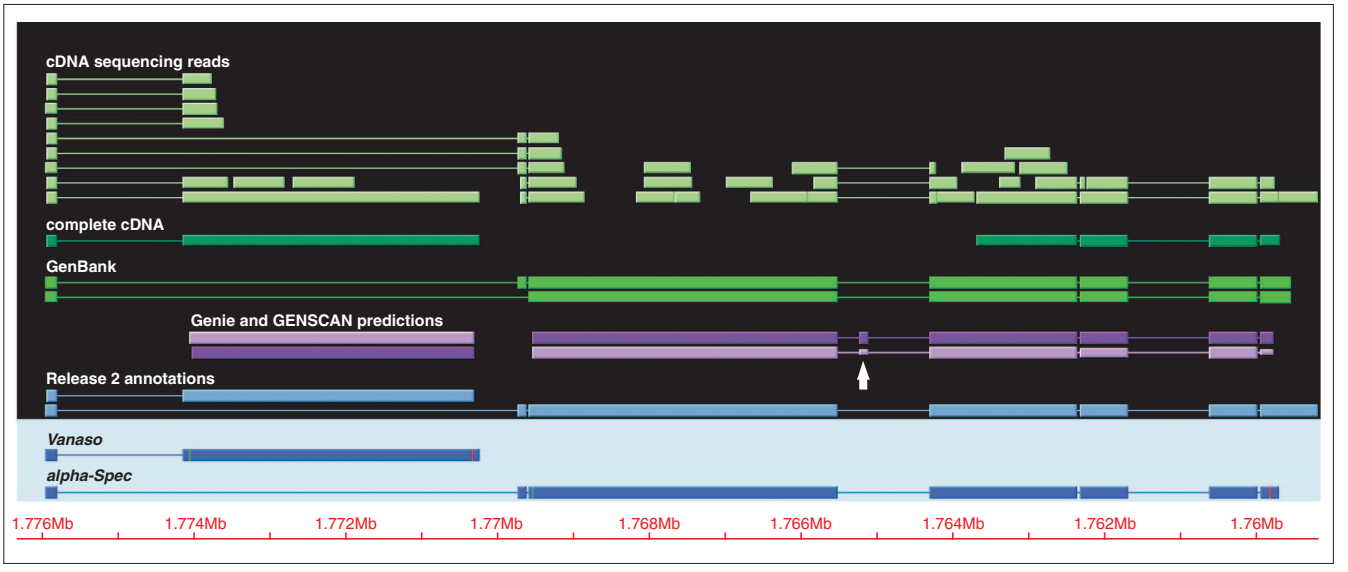

Figure 7

Vanaso and *α-Spec* are separate annotations that share an untranslated 5′ exon. Only evidence (black panel) directly used to annotate the gene models (cyan panel) is shown. Coding sequences are delineated by green vertical lines (starts of translation) and red vertical lines (stops of translation). The Release 3 annotations *Vanaso* and *α-Spec* (dark blue) on chromosome arm 3L overlap at their most distal 5′ end, sharing a portion of their untranslated regions. These gene models are supported by many ESTs and cDNA sequencing reads (light green), a complete cDNA clone (dark green), and several GenBank records (dark green). In spite of the shared initiation point for these transcripts, none of the remaining exons or coding sequences coincides. Note the small exon (arrow) predicted by Genie and GENSCAN. This exon is not included in the *α-Spec* annotation, for lack of other supporting evidence, but alternative cDNA clones including this exon will be screened for directly in cDNA libraries [30].

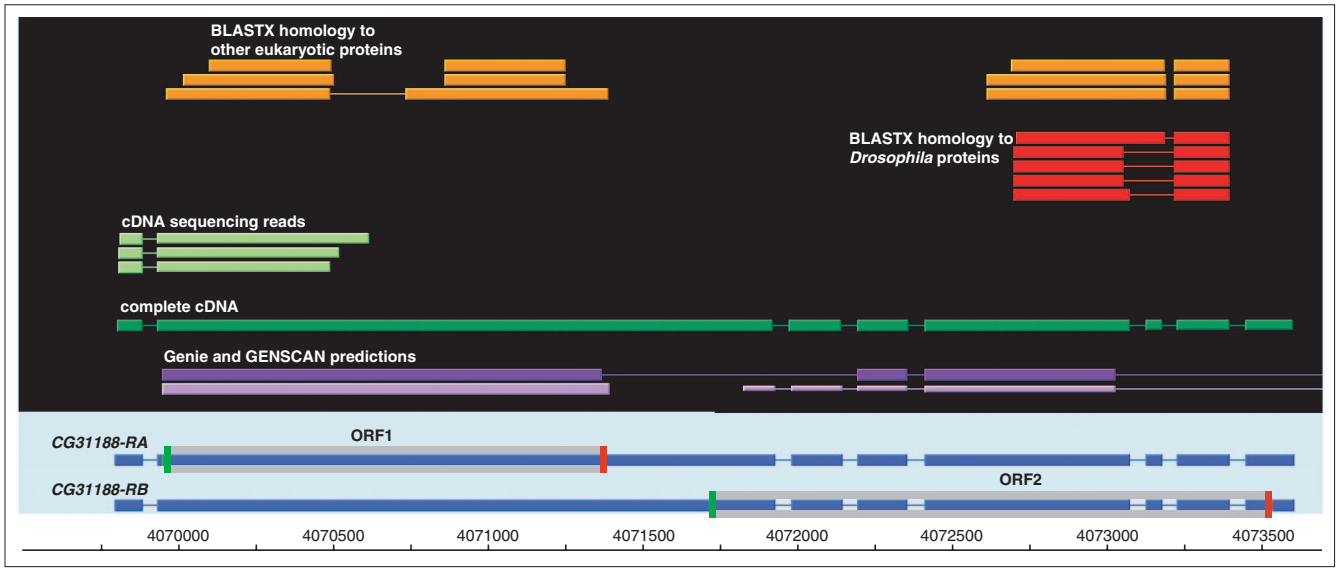

Figure 8
CG31188 is a dicistronic gene. Data directly used to annotate the dicistronic gene model are shown in the black panel and the gene models generated from these data are shown in the cyan panel. Coding sequences are delineated by green vertical lines (starts of translation) and red vertical lines (stops of translation). Dicistronic genes (dark blue) were predicted when assembled cDNA sequencing reads or complete cDNA sequence (light and dark green) span two complete open reading frames (ORF1 and ORF2, shaded in cyan panel) that are separated by in-frame stop codons. There must be additional evidence supporting the existence of both predicted peptides. In the case of *CG31188* on chromosome arm 3R, each of the two ORFs shares homology with proteins from other eukaryotes (orange) or *Drosophila* (red).

regions and the number of ATGs in these regions, and found that there are seven pairs of dicistronic genes that appear to conform to this pattern (indicated in Supplementary Table 3). The majority of dicistronic cases clearly do not conform to such a model of partial ribosome disassembly and continued scanning. In four cases, the intercistronic region is less than 4 bp. However, most of the annotated dicistronic pairs are separated by several hundred base pairs, and the separation can be as much as 4.5 kb. In these longer intercistronic regions, there may be multiple ATG codons before the second translation start site. A mechanism of translation initiation utilizing an internal ribosome entry site (IRES [63]) appears a better explanation for these cases. Oh *et al.* [64] have hypothesized that certain *Drosophila* genes with long 5′ UTRs might be translated via internal ribosome entry. If this is the case, translation of the second CDS within a dicistronic transcript may be effected by the same initiation mechanism.

Assessment of Release 3 quality

Did we miss genes?

Andrews *et al.* [16] suggested that the number of genes in *Drosophila* might be a severe underestimate, based on 7,297 testes EST sequences they generated and aligned to the annotated genome. However, using their data, as well as 23,087 additional testes-derived ESTs [29], we predict a similar number of genes in Release 3 as in previous releases. The more likely explanation for their results is that 5′ exons, and not genes, were under-predicted in *Drosophila*, since there is EST evidence for testes-specific promoters and transcripts [29]. In most cases, the testes ESTs did not align to Release 1 genes because only the downstream CDS had previously been annotated for those genes, whereas in Release 3 the UTRs that match the testes ESTs are annotated.

Gopal *et al.* [15] reported 1,042 novel genes that were not included in the Release 2 annotation. After completing our reannotation, we compared this set of 1,042 genes to our data. We found that 75% of their predicted genes mapped to euchromatin, 16% mapped to heterochromatin, 7% mapped within transposable elements, and 1% could not be found in the Release 3 genomic sequence. Of the 75% that mapped to euchromatin, 66% (520) do match Release 3 annotations. The remaining 34% did not match Release 3 annotations, leading to the possibility that some or all of these may represent novel genes. Incorporating their methods (threading GENSCAN predictions to look for structural homology) into our computational approach may uncover additional missed genes.

One way to address the quality of the Release 3 annotations is by comparative sequence analysis. In an accompanying paper, C. Bergman *et al.* [65] surveyed sequence conservation in approximately 0.5 Mb of the *D. melanogaster* genome containing 81 genes using comparative data from four *Drosophila* species (*D. erecta, D. pseudoobscura, D. willistoni* and *D. littoralis*). Their comparison to our *D. melanogaster* annotations detected no genes conserved in other species that were missed in Release 3 [65].

Other genes likely to be missed are genes with small ORFs (because of the arbitrary length cutoffs we used, see Materials and methods), and genes expressed transiently during development, at very low levels, and/or in cells and tissues not represented by the DGC cDNA libraries. Future DGC cDNA clones will be generated by directed screening of cDNA libraries with probes matching predicted exons [30], and cDNAs selected during the re-annotation process to represent alternative transcripts not currently in the DGC.

Reliance on gene prediction data versus cDNA data

Of the final set of annotations, 93% contain sequences that are present in Genie-predicted exons and 96% contain sequences that are present in GENSCAN-predicted exons (Table 3). Only 249 (2%) of protein-coding gene models were created without an *ab initio* model, that is, solely on the basis of cDNA or protein homology evidence. The fact that 98% of our accepted annotations span a region containing a gene prediction supports both the strength of the prediction programs' algorithms as well as our reliance on them for our methods. However, both Genie and GENSCAN gene models were often wrong in detail, when compared to cDNA sequence alignments for three main reasons: first, exon misassociations: the programs placed exons from one gene with the exons of a neighboring or nested gene; second, erroneous splice site calls: the donor and acceptor sites were slightly misplaced; or third, missed mini- and micro-exons: small ORFs were not identified [20,66].

The fraction of gene models that are based solely on gene prediction data has decreased considerably, from 2,348 (17%) in Release 1 to 815 (6%) in Release 3 (Table 3). This shift was primarily due to the more recently available *Drosophila* EST and cDNA sequences, rather than newly evident similarity to sequences in other species.

Alignment of full-length cDNA sequences from the same strain continues to be the best way to annotate gene models [66-68]. The number of ESTs generated by the BDGP project increased from around 86,000 in Release 2 to 246,248 in Release 3 [29] and the number of sequenced full-insert cDNAs from around 1,000 in Release 2 to over 9,000 in Release 3 [30]. For approximately 6,000 of these, the completely assembled sequence was available during the re-annotation effort (see Materials and methods). In addition, 8,699 ESTs from the community deposited in dbEST [69], including a set of 7,297 from a testes cDNA library [16], were available. In all, 78% of the protein-coding genes show a match to an EST sequence (Table 3) and over half to full-insert cDNA sequences. We anticipate further improvement to gene models as more cDNA data become available.

Non-consensus splice sites and small introns

All introns within protein-coding genes were examined for conserved GT/AG splice junctions with the Sequin program [70,71], and all instances of annotations lacking GT/AG splice junctions were inspected and commented on. Of the 48,039 total splice junctions, 0.5% are annotated with GC/AG splice junctions, a frequency that might justify describing GC as an alternative splice donor. Eleven instances of AT/AC splice junctions are annotated. An especially well supported example of AT/AC usage is *CG1354*, which has more than 25 confirming ESTs. Other cases of non-consensus splice sites appear rare; however, more are likely to be documented in the future. The particular alignment algorithm used (see Materials and methods) and our reliance upon gene-prediction data imposed a bias against unconventional splice sites. In a number of cases for which an unconventional splice junction was supported by cDNA data, the precise location of the junction could not be determined, owing to repeated sequence at the donor and acceptor sites. A good example of this type of pattern is *sba (CG13598)*. Two alternative transcripts for this gene are supported by cDNA data, and both appear to contain an unconventional, ambiguous splice junction. The two transcripts share the unconventional splice acceptor site; they differ in the location of the non-consensus splice donor site, but the two donor sites are identical in sequence.

We also examined every gene model with an intron less than 48 bp. The frequency of such introns in *Drosophila* is low; 32 are annotated in Release 3. There are several well supported examples of introns less than 45 bp, with at least two supporting cDNAs derived from the sequenced strain. These include *mod(r)* and *csul*, each with an intron of 44 bp, and *CG11892*, with a diminutive intron of 43 bp.

SWISS-PROT/TrEMBL validation of the models

We used the SWISS-PROT and TrEMBL protein databases [13] and the PEP-QC software program [12] to validate the integrity of the annotations and to track consistency with previously published data (see Materials and methods). Of the 3,687 annotated peptides with a cognate in the curated SWISS-PROT/TrEMBL dataset, 75% (2,764) were of identical length and had more than 99% sequence identity. Curators examined each case with less than 100% sequence identity, and in some cases, annotation errors were detected and corrected. For example, translation start sites were shifted to the experimentally reported position, which in some cases was downstream of the predicted start. However, in most cases discrepancies appeared to be due to strain-specific polymorphisms or errors in the reported DNA sequence on which the SPTRreal entries were based (see Materials and methods). Given the high quality of the underlying Release 3 genomic sequence, we believe that in many cases the Release 3 annotation is more accurate than the sequences deposited by the community in SWISS-PROT and TrEMBL.

Confidence in Release 3 gene models

The amount of evidence attributable to each Release 3 gene model varies considerably, and therefore our confidence in these gene models, even the confidence in two alternative

transcripts encoded by the same gene, may differ greatly. To estimate the reliability of a gene model, we developed a classification system that groups data into four categories: computational gene predictions; protein similarities; alignments of ESTs and other partial cDNA sequences; and alignments of full-insert cDNA sequences. One point was assigned to each data type that overlapped a given annotation, and a score of 1 to 4 was determined for each transcript, with 1 being the lowest and 4 being the highest confidence (see Materials and methods). As shown in Table 4, more than 80% of transcripts and more than 75% of genes were assigned a confidence value of 3 or 4. Thus, we have high confidence in a large proportion of the Release 3 gene models.

Limitations in our methods

The Release 3 annotations should be much more consistent than in previous releases because fewer curators were involved, a defined set of rules was used, and additional validation steps were performed (see Materials and methods). We set a rigorous standard for the annotations by requiring attributable evidence for every gene model, for example, a gene prediction, an alignment to a GenBank/EMBL/DDBJ accession, or a curated personal communication to FlyBase. However, during the Release 1 analysis, *Drosophila* researchers annotated particular families of genes about which they were expert and for which they may have had specific unpublished information. Much of the evidence for these annotations was not released to the public domain and is not currently available, so some of the details of these gene models were lost in Release 3. The solution in these cases is for biologists in the community to continue to submit error reports to FlyBase to be curated by FlyBase as personal communications. The resulting set of annotations will be stronger because every gene has traceable evidence that is available in the database and is annotated according to a standard set of rules.

As is expected with such a complex analysis, rules cannot be expected to cover every eventuality. As a result, some of the annotations are based partially on curator judgment, introducing a potential source of inconsistency. Visual inspection and curator expertise were absolutely necessary in overcoming shortcomings of the automated processes such as identifying GC splice donors and sorting out complex gene models. It was also essential for annotating unusual cases, such as the dicistronic genes and overlapping gene models. Further, it should be noted that manual annotation is an iterative process. Subsequent to an initial annotation call, a set of automatic verification steps was carried out. Potential errors were reviewed and, where appropriate, annotations were modified as a result of the verification analysis.

Accessing data and reporting errors

The Release 3 genomic sequence available at GenBank/EMBL/DDBJ [7] includes all gene models, that is, the extent of transcripts and each corresponding CDS. More complete information, including all classes of evidence, can be obtained from FlyBase, presented in Gadfly Gene Annotation reports, in interactive Genome Browser maps, in the Apollo annotation tool, and by batch download. In addition to transcript structures, the Gene Annotation report presents the evidence supporting a gene model, any comments included by the annotator, and a thumbnail view of the immediate genomic region. There are links to the reports for adjacent genes, to the FlyBase Genome Browser view of the surrounding region, and to FASTA files of protein, transcript, and genomic sequences. Another link takes users to the results of automated BLASTP and InterProScan [72] analyses of the predicted peptides. The coordinates, comments, and sub-features of the annotations (such as UTRs, exons, and so on) can be downloaded in a number of formats, including XML and GFF. The interactive genome browser shows all transcripts annotated within a region; a zoom feature allows the user to choose the level of resolution. Additional data classes can be added, at the discretion of the user, including the extent of DGC cDNA clones and EST data, the BAC clones used for determination of the genomic sequence, and the position of P-element insertions isolated by the BDGP Gene Disruption Project [73].

Researchers can also use the Apollo genome annotation and curation tool [19] to view the supporting data in greater detail. This Java software tool is available for local installation [74] and bulk downloads of the annotations and computational evidence are available in XML or GFF format [75]. Sequence data in multiple FASTA format for the entire set of annotations are also available at this site. In addition, Apollo includes software to request and retrieve the annotations and other data transparently from FlyBase/BDGP. Many individual investigators have already contributed substantially to the Release 3 annotations by submitting corrections to gene structures using the error report forms [76], and researchers can continue to submit reports to FlyBase in this manner. In the future, we hope that by enabling researchers to send an Apollo XML output file to FlyBase for review, error reporting of fine gene structures will be simplified.

Future updates

Changes to the sequence

The BDGP will continue to finish the remaining problematic regions of the euchromatic genomic sequence to high quality (see [17]), and focus efforts on refining the sequence of the heterochromatin [25]. Changes to the sequence will be submitted to GenBank/EMBL/DDBJ every 6 to 12 months.

Because sequence updates at the time of new releases will result in changes to the coordinate system for each chromosome arm and for GenBank/EMBL/DDBJ accession units, it will be particularly important for researchers to make note of specific release and version dates when providing sequence coordinates. FlyBase encourages researchers to

refer to coordinates as associated with specific GenBank/EMBL/DDBJ accession and version numbers.

Changes to gene models

Future re-annotation will be on a gene-by-gene basis, rather than a survey of the entire genome. Future analyses will include new large-scale datasets, including the *Anopheles gambiae* genomic sequence, the *D. pseudoobscura* genomic sequence, and additional DGC cDNA sequences [30]. Changes to the gene models will occur more often than changes in the sequence, and will be reported in date-stamped updates of the GenBank/EMBL/DDBJ accessions and FlyBase records. Such changes are reflected in the feature annotations only and thus do not constitute new releases, as the underlying genomic sequence does not change.

FlyBase will also focus on the localization of many more annotation features to the genome view, such as regulatory elements, mutational lesions, rearrangement breakpoints, and P-element insertion sites. Many of these sequence features are already in the FlyBase genetic data tables and gene annotation reports, based on data from literature curation, computational analyses (for example [77]), and large-scale projects such as the BDGP Gene Disruption Project.

Changes to functional annotation

In the annotation of genes in Release 1 attributes of gene products were predicted with respect to their molecular functions, the roles they might play in biological processes, and their cellular locations, using the controlled vocabularies developed by the Gene Ontology (GO) Consortium [78]. These predictions were computational, using a program known as LOVEATFIRSTSITE written by M. Yandell [3]. Since then, FlyBase curators have assessed each of these annotations, retaining in FlyBase only those that were reasonably secure, and have re-annotated many genes with GO terms of higher granularity. This work, together with the curation of GO terms from the literature and sequence records, has resulted in 7,299 genes sharing 25,057 GO annotations. This analysis has not yet been repeated for the Release 3 gene products, but the curation of GO terms for all new genes and all split/merged genes is now in progress. When this annotation is completed we will have a benchmark for further automatic predictions of GO terms, using programs similar to LOVEATFIRSTSITE [3] and PANTHER [79].

Conclusions

Annotation of eukaryotic genomes is not a straightforward process, owing to the limitations of the current gene-prediction algorithms. However, we have made the annotation process much more rigorous by utilizing a large set of experimental data, manual curation, and defined standards. By using a large amount of cDNA alignment data and a tool facilitating the rapid visual inspection of evidence for each gene model, we were able to significantly improve the quality of *Drosophila* gene annotations. We found that a comprehensive set of curation rules was crucial to making manual annotation consistent and reliable. We also found that comparison of predicted peptides to experimentally verified SWISS-PROT and TrEMBL sequences was an important quality-assessment step. In future, we plan to make the automated analysis of predicted polypeptides, including identification of their protein domains and sequence similarities, a more integrated part of genomic sequence annotation. Finally, by making the annotations, comments, and all supporting evidence available to users, we have provided the scientific community with the resources to assess the quality of each gene model.

Our analysis reveals a number of genes that fall outside the definition of conventional gene models: neighboring genes with overlapping UTRs; genes with alternative transcripts encoding distinct coding regions; and dicistronic transcripts. An even larger number of genes show alternative splicing or are nested within neighboring genes. Currently, gene-prediction algorithms are unable to accurately predict such gene models. Studies like this one are a prerequisite to extending current computational methods to more successfully and specifically predict eukaryotic gene structures, by defining the classes of features and the requirements for supporting evidence. Once sophisticated computational pipelines can cope with the full range of complex genomic features, we will benefit from better resources for biological investigation.

FlyBase is one of several major model organism databases with high-quality euchromatic sequence charged with curation of experimental data from the literature. Unlike many other organisms, *Drosophila* has a genetic history reaching back to 1910, and an enormous amount of data to tie to the sequence. In this paper, we have addressed one of the first challenges, accurately annotating the genomic sequence, by utilizing the extensive resource of full-insert *D. melanogaster* cDNA sequences and FlyBase gene records (containing existing community data), and by manually curating the gene models using defined methods and controlled vocabularies. However, there is more work necessary to tie the annotated genomic sequence and annotated peptide sequences to further experimental data from the literature, results of large-scale analyses (for example, microarray expression data), and new computational analyses (for example, comparative sequence analysis). We believe shared data-exchange formats and ontologies will be vitally important to curate, collate, and structure this huge amount of data in a way that allows researchers to exploit the information to its full potential.

Materials and methods

Re-annotation of the euchromatic genome was performed by dividing the long finished chromosome arm sequences from the BDGP into 250-350 kb segments roughly corresponding

to the Release 2 sequences available at GenBank/EMBL/DDBJ [6-11], running a 'pipeline' of computational analysis steps on this sequence [12], and allowing one curator to annotate all of the genes on one segment using the genomic feature editor, Apollo. Apollo is a new graphical user interface developed in a collaboration between FlyBase-BDGP and Ensembl, that allows curators to view the results of computational analyses and to edit the annotations efficiently [19]. Curators manually examined 437 segments, constituting 117 Mb of euchromatic sequence. We note that because of sequence finishing and other adjustments, the length, composition, and end sequences of some updated Release 3 submissions may not match the Release 2 submissions, but most of the genes remained on the accession in which they were annotated in Release 2.

We aligned to the genomic sequence 254,947 *Drosophila* ESTs and over 9,000 full-insert cDNA sequences from the BDGP [29,30] and the community. We also incorporated protein data from BLASTX sequence similarity searches [22,23] of the SWALL (SWISS-PROT/TrEMBL/TrEMBLNEW) peptide dataset [13,80,81] from a broad range of species.

Curation rules

We have attempted to provide documentation for as many annotation decisions as possible. In addition to providing access to evidence (EST and full-insert cDNA sequence reads, prior sequence submissions, BLASTX homologies, and gene prediction data), we have developed and made available a set of annotation rules (see [82] and additional data files) and have provided textual comments to explain atypical or subjective annotations.

The annotation rules promote consistency in the annotation effort, and deal with all aspects of annotation: from assessment of whether a marginal gene prediction should be the basis for a new gene model to the annotation of atypical splice sites; from the determination of alternative transcriptional starts and stops, and the designation of translation starts, to the use of comments to flag atypical or questionable annotations. Cases with insufficient, atypical, or conflicting data that the rules did not address were left to the discretion of the annotator; in such instances, comments to document the subjective nature of the gene model were added.

Typically, at least one annotation was created containing each site of alternative splicing represented in the EST/cDNA data. For atypical splice junctions, a higher level of supporting data was required (see below). Often, sites of alternative splicing were supported by ESTs but not full-insert cDNAs. Since, as a matter of policy, we tried to avoid creating partial transcript models, this required that we postulate transcripts combining, for example, 5′ and 3′ ESTs corresponding to different cDNAs. In some cases, these combinations may not exist *in vivo*. In particularly complex cases, curators did not create every splice form suggested by the data, but commented that the potential exists for additional splice forms.

The rules used were specific for this annotation effort, in particular, for the types of data currently available. For example, because of the limited amount of 3′ EST data, little attempt was made to annotate alternative transcripts that differ as a result of multiple polyadenylation sites.

Establishment of the annotation rules included the development of a set of controlled comments, that is, comments that are reproducibly phrased and are consistently used. Such controlled comments were used to confirm atypical gene structures, such as the use of atypical splice sites or overlapping UTRs, and to document the evidence used in subjective cases, such as an unusual gene structure based on a single EST or a gene model based solely on gene-prediction data. DGC cDNA clones that appeared to contradict other evidence were also flagged; most frequently, these were not full-length or appeared to contain intronic sequences.

Annotation of non-protein-coding genes

To annotate small, non-protein-coding RNA genes previously collected in the FlyBase database, we retrieved sequences for each gene from GenBank [6,7] and generated a multiple-FASTA dataset. Occasionally, sequence was retrieved from the original literature. The FASTA dataset was then aligned to the Release 3 genome by Sim4 alignment. MicroRNAs were aligned by BLASTN analysis; a single exact match was found for each of the microRNAs listed in FlyBase.

Evidence for gene structures

Gene prediction data

The publicly available version of Genie, which does not utilize EST or BLAST evidence [20], predicted 13,794 genes on the finished sequence. GENSCAN predicted 19,189 genes. As reported previously [3,20], Genie appears to predict fewer false positives, perhaps because it has been trained on *Drosophila* sequences, whereas GENSCAN has only been trained on vertebrate datasets [21]. However, GENSCAN also shows greater sensitivity than Genie, identifying some real genes that Genie fails to find. To balance the false-positive and false-negative rates of GENSCAN, we used an empirical prediction score as a threshold, as done previously [44]. In the absence of other supporting evidence for a gene, we used GENSCAN predictions only when at least one exon had a score > 45; this is a stringent threshold, as 21% of the genes in Release 3 with full-length cDNA evidence do not contain any exons scoring > 45.

BLASTX/TBLASTX sequence similarity data

To detect proteins with significant sequence similarity, we used BLASTX to compare translated genomic sequence to peptides in other species included in SWALL [13,83], and

TBLASTX to compare translated genomic sequence to virtual translations of the rodent UniGene set [84] and insect sequences in dbEST [69]. We also looked for sequence similarities to *Drosophila* peptides that had experimental verification, but not to those representing purely hypothetical or computational gene models (see below). Although the number of proteins in a public database like TrEMBL [13] has increased exponentially in the time between the Release 1 annotation in November 1999 and the Release 3 annotation in 2002 [85], the increased size of the protein datasets resulted in a 14% increase in the number of fly genes that produce proteins with similarity to other proteins. In March 2000, Adams *et al.* [3] reported that 9,612 (71%) of the 13,601 of the Release 1 genes showed a match to another protein. We now find that 10,996 (82%) of the Release 3 protein-coding genes show a match by BLASTX or TBLASTX (with expectation value less than or equal to $1 \times e^{-7}$). However, we note that the datasets we used were fixed before the release of the genomic sequence of *A. gambiae* [86], the only other dipteran (or arthropod) with a complete genome sequence. We expect that a higher percentage of *Drosophila* proteins will show sequence similarity to *Anopheles* proteins, because *A. gambiae* is more closely related to *D. melanogaster* than are the other available model organisms [86].

EST and cDNA alignment data

Prediction of gene models was made more rigorous by the increased availability of cDNA data. However, misleading alignments can be created by the presence of genomic DNA contaminants, cDNA clones containing two independent cDNAs co-ligated in the same plasmid vector (chimeras), and internal priming of cDNAs during library synthesis. cDNA clones derived from incompletely processed primary transcripts are not readily distinguishable from alternative splicing without experimental verification. Moreover, cDNA sequences designated as full length may actually be truncated; approximately 1,000 of the 9,000 full-insert sequences from the BDGP are probably not full-length [29,30]. The Sim4 alignment tool can make mistakes in determining splice site junctions or completely fail to align very small exons [24,67]; indeed, a small number of cases of failure to align microexons were identified by Stapleton *et al.* [30] when they compared the predicted translation products of cDNAs with those of Release 3 gene models. However, Haas *et al.* found est2genome and other alignment tools were, in general, not superior to Sim4 [24,67]. Despite these limitations, alignment of complete cDNA sequences is invaluable in detecting UTRs, alternative splicing events, detailed exon-intron structures, nested genes, and other key aspects of gene models.

Full-insert *Drosophila melanogaster* cDNA sequences came from a number of sources. The largest set of full-insert cDNA sequences came from the BDGP *Drosophila* Gene Collection (DGC) project [5,29,30]. Of the protein-coding genes, 9,297 (69%) show a match to full-insert sequences from the cDNA clones in the DGC, and in some cases, more than one DGC clone provided definitive gene models for alternatively spliced products. At the time of annotation, we had access to full-insert sequencing reads from 9,074 of the 10,910 cDNA clones, but only some 6,000 of these had been fully assembled. Gene models based on incompletely assembled cDNA clones were marked 'incomplete'. These gene models will be among the first annotations to be updated.

Sequences deposited in public databanks like GenBank/EMBL/DDBJ [6-11] by *Drosophila* researchers provided definitive evidence for a number of genes. For a subset of well-studied genes, FlyBase curators synthesized all of the available sequence and literature data into high quality Annotated Reference Gene Sequences (ARGS) that have been deposited in GenBank's RefSeq division [1]. These ARGS sequences correspond to 795 (6%) of the Release 3 annotations.

Other sequences came directly to FlyBase as error reports from the scientific community. FlyBase curated 636 reports with information about 1,094 genes as personal communications, and any sequences supplied in these reports were aligned to the genome. In all, 825 (6%) of the annotations overlapped these sequences (Table 3). Accurate annotation of three gene families in particular was greatly facilitated by sequence submitted in error reports: 85 cytochrome P450 monooxygenase genes (B. Dunkov, personal communication, FBrf0132129, FBrf0126925; D.R. Nelson, personal communication, FBrf0136021), 80 gustatory receptor genes (H. Robertson, personal communication, FBrf0141780; K. Scott and R. Axel, personal communication, FBrf0137428), and 61 odorant receptor genes (H. Robertson, personal communication FBrf0136024; C. Warr and L. Vosshall, personal communication, FBrf0128191).

Determination of confidence values

The extent of each transcript and corresponding CDS was extracted from the '*Drosophila* Genomic Sequence Annotations' file (in GFF format [87]), which is available [75]. The extent of overlap of each transcript against the supporting evidence used during the re-annotation was determined using an intersection algorithm to determine the annotations overlapped by particular types of evidence [12].

The evidence datasets used included: gene prediction based on Genie [20] and GENSCAN [21]; Sim4 alignments to EST and full-insert cDNA sequencing reads derived from the BDGP cDNA project [29,30], the earlier analysis of the *Adh* region [44], and dbEST (for example [16]); FlyBase ARGS [1]; GenBank/EMBL/DDBJ entries identified as *Drosophila* cDNA sequences [6-11] and error report submissions to FlyBase [1,2]; and BLASTX protein homology data. For a complete list of the evidence datasets and their description see [82]. Data were filtered using the Bioinformatics Output Parser (BOP), which also assembled all EST and full-insert cDNA sequence reads

from a particular cDNA clone into a virtual assembly (BOP [12]). The Apollo tool displayed these assemblies with sequence gaps indicated differently from introns.

The algorithm used to assess relative annotation quality assigned one point for overlap of a gene prediction, either Genie or GENSCAN or both. One additional point was assigned for overlap with protein similarity data. The remaining datasets were considered in the following order and resulted in an additional one or two points: the cDNA and annotation data were analyzed to determine if any entry in this class spanned the entire length of the CDS; if so, an additional two points were assigned, and if not, GenBank/EMBL/DDBJ and error report entries were analyzed and if any spanned the length of CDS, two points were assigned; if none of these data classes corresponded to the full-length CDS, then the existence of partial cDNA data and/or overlapping EST data merited one point. Details of the rules for this classification system can be found at [82].

Integrity checks

SWISS-PROT/TrEMBL validation of the translated models by PEP-QC is described below. Both annotated segments and chromosome arms were validated using the Sequin software tool from the NCBI [71], which found mistakes in exon-intron structure, start of translation, and ID duplication. We queried our dataset for proteins < 50 amino acids, CDS features making up less than 25% of the predicted transcript length, introns < 48 bp, and visually inspected each annotation in these classes making comments where appropriate. We checked annotations that overlapped transposable elements and tRNA genes, or appeared multiple times in the genome with duplicate identifiers. We verified that deleted Release 2 annotations had no independent evidence in literature-curated references. Finally, in order to allow the construction of a wild-type proteome from the mutant sequenced y^1; cn^1 bw^1 sp^1 strain, we replaced annotated sequences from known mutated genes (*y, cn, bw, MstProx, LysE, Rh6*) with RefSeq wild-type sequences from GenBank with an appropriate note.

Non-consensus splice sites

All introns within protein-coding genes were examined for conserved GT/AG splice junctions with the Sequin program [70,71], and all instances of annotations lacking GT/AG splice junctions were inspected and commented upon. Splice junctions were based upon alignment of cDNA/EST sequence and, in the absence of such data, on gene prediction models. Even for transcript structures based upon EST data, the number of atypical splice junctions is probably an underestimate. The alignment algorithm used (Sim4) forced intron junctions to occur at GT/AG sites whenever possible, even at the expense of a several-base mismatch. This occasionally resulted in apparent early translation termination, in which case the annotator checked for a GC donor that would allow read-through. Other GC splice annotations were based on information in the literature or GenBank/EMBL/DDBJ records. With the exception of GT/AG junctions, we imposed a higher standard of verification for unconventional splice annotations: sequence data from a cDNA isolated from the sequenced strain, or multiple consistent ESTs.

SWISS-PROT/TrEMBL validation of the models

The SWISS-PROT and TrEMBL protein databases [13] were used to validate the integrity of the annotations and to track consistency with previously published data. The SWISS-PROT Protein Knowledgebase [80] is a curated protein sequence database that provides a high level of annotation, a minimal level of redundancy and high level of integration with other databases. The TrEMBL database [81] contains the translations of all CDS present in the EMBL Nucleotide Sequence Database [8,9], which are not yet integrated into SWISS-PROT [80]. A non-redundant set of SWISS-PROT and TrEMBL *Drosophila* sequences was created, and sequences representing purely hypothetical or computational gene models (those corresponding to CG, BG, or EG genes in FlyBase) were excluded. The PEP-QC program [12] compared the resulting collection of 3687 *D. melanogaster* sequences (SPTRreal) to the annotated peptides using BLASTP [22]. Each gene was placed into one of four 'validation' categories: Perfect match to SPTRreal (annotated peptide of identical length with 100% sequence identity), Single AA substitutions (annotated peptide of identical length with ≥ 99% sequence identity), Significant mismatch (annotated peptide and SPTRreal entry do not align over their entire length, but do contain aligned spans of 40 residues or 20% peptide length, with at least 97% sequence identity), or Poor match (poor or no BLAST hits). For the first two categories, the SPTRreal peptide was allowed to match a portion of the annotated peptide if it was designated as a 'fragment' of the full peptide sequence (125 or 3.4% of SPTRreal genes are designated with this tag).

In Release 3, about 75% of all genes with a cognate in SPTRreal are in the Perfect match and Single AA substitution categories (2,764 out of 3,687). Curators examined every gene marked Single AA substitutions, Significant mismatch, or Poor match. In some cases, annotation errors were detected and gene models were modified to produce a translation matching the SPTRreal sequence. For example, curators used the PEP-QC output to change translation start sites to the experimentally reported position where appropriate. In other cases, discrepancies may be due to strain polymorphisms, errors in the reported DNA sequence on which the SPTRreal entries were based, or undetected errors in either the Release 3 annotation or sequence. Given the high quality of the underlying Release 3 genomic sequence, we believe that in many cases the Release 3 annotations are more accurate than the SPTRreal sequences.

The improvement over the Release 2 annotations is evident. The combined number of Perfect match and Single AA

substitution annotations has increased by 22%, from 2,260 in Release 2 to 2,764 in Release 3, while the combined number of Significant mismatch and Poor match annotations has decreased by 35% (1,427 in Release 2 versus 923 in Release 3). Over the past two years, SWISS-PROT has incorporated sequences from Releases 1 and 2 of the genomic annotations into the curated protein sequences (E. Whitfield, personal communication). Therefore, while the overall quality of the sequences in SWISS-PROT has almost certainly increased during this time, it is still worth asking the question - how well do the Release 3 annotations match SWISS-PROT sequences deposited before the initial genome annotations? This question was answered for the Release 1 annotations using a dataset of 1,049 polypeptide sequences created in SWISS-PROT before 1999, prior to the publication of the annotated Release 1 sequence [14]. That group found that 578, or 55%, of this SWISS-PROT set had an annotated peptide of identical length with at least 95% amino acid identity (A. J. Gentles, personal communication). Performing the same analysis using the Release 3 annotations, we find that 694, or 66%, of the peptides in the same SWISS-PROT dataset match with 95% identity. This 20% increase in the number of matching polypeptides most likely is a reflection of the improvement in the quality of the annotations. But what about the 34% of cases that show significant mismatch between Release 3 and SWISS-PROT? We have examined these cases and believe that the large majority of these remaining discrepancies are not due to mistakes in our annotation, but due to strain polymorphisms, as well as errors in the sequence underlying the SWISS-PROT entries created before 1999. If this were indeed the case, then the analysis of Karlin *et al.* [14] would have grossly overestimated the error rate in the Release 1 annotations.

Drosophila *genes*

References for all *Drosophila* genes mentioned in the paper can be found in FlyBase [1,2].

Additional data files

Supplementary tables of annotated pseudogenes, the distribution of alternatively spliced transcripts, and dicistronic genes, along with the guidelines for re-annotation, are available with the online version of this paper.

Acknowledgements

This work was supported by NIH grant HG00750 to G.M.R., by NIH Grant HG00739 to FlyBase (W.M.G.), by MRC Grant GB9827766 to FlyBase (M.A.), and by the Howard Hughes Medical Institute (C.J.M., J.R., and G.M.R.). We would like to thank C.R. Nelson, C.M. Bergman, A. Page-McCaw, and R. Hoskins for helpful discussions, B. Kronmiller, J.W. Carlson, D. Emmert, B. Marshall, F. Smutniak, C. Wiel, and M. Yandell for expert assistance with datasets and computational analysis, M.E. Clamp and S.M.J. Searle for their work on the Apollo annotation tool, S. Mount for help with the snRNA dataset, and E. Frise and E. Smith for their technical support. We also thank the authors of the FlyBase error reports, in particular H. Robertson, T. Benos, B. Dunkov, D.R. Nelson, K. Scott, R. Axel, C. Warr, and L. Vosshall; K. Burtis, S. Langley and J. Carlson, for communicating unpublished data; and A. Gentles, S. Karlin, and T. Gaasterland for providing us with the data used in their 2001 publications. Finally, we wish to thank C.M. Bergman, R. Hoskins, A. Huang, and C. Siebel for their critical reading of the manuscript.

References

1. The FlyBase Consortium: **The FlyBase database of the *Drosophila* genome projects and community literature.** *Nucleic Acids Res* 2002, **30:**106-108.
2. **FlyBase: a database of the *Drosophila* Genome** [http://www.flybase.org]
3. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, *et al.*: **The genome sequence of *Drosophila melanogaster.*** *Science* 2000, **287:**2185-2195.
4. Rubin GM, Yandell MD, Wortman JR, Gabor Miklos GL, Nelson CR, Hariharan IK, Fortini ME, Li PW, Apweiler R, Fleischmann W, *et al.*: **Comparative genomics of the eukaryotes.** *Science* 2000, **287:**2204-2215.
5. Rubin GM, Hong L, Brokstein P, Evans-Holm M, Frise E, Stapleton M, Harvey DA: **A *Drosophila* complementary DNA resource.** *Science* 2000, **287:**2222-2224.
6. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Rapp BA, Wheeler DL: **GenBank.** *Nucleic Acids Res* 2002, **30:**17-20.
7. **GenBank via ftp** [http://www.ncbi.nlm.nih.gov/Genbank/GenBankFtp.html]
8. Stoesser G, Baker W, van den Broek A, Camon E, Garcia-Pastor M, Kanz C, Kulikova T, Leinonen R, Lin Q, Lombard V, *et al.*: **The EMBL nucleotide sequence database.** *Nucleic Acids Res* 2002, **30:**21-26.
9. **EMBL nucleotide sequence database** [http://www.ebi.ac.uk/embl]
10. Tateno Y, Imanishi T, Miyazaki S, Fukami-Kobayashi K, Saitou N, Sugawara H, Gojobori T: **DNA Data Bank of Japan (DDBJ) for genome scale research in life science.** *Nucleic Acids Res* 2002, **30:**27-30.
11. **DDBJ DNA Databank of Japan** [http://www.ddbj.nig.ac.jp]
12. Mungall CJ, Misra S, Berman BP, Carlson J, Frise E, Harris N, Marshall B, Shu S, Kaminker JS, Prochnik SE, *et al.*: **An integrated computational pipeline and database to support whole-genome sequence annotation.** *Genome Biol* 2002, **3**:research0081.1-0081.11.
13. Bairoch A, Apweiler R: **The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000.** *Nucleic Acids Res* 2000, **28:**45-48.
14. Karlin S, Bergman A, Gentles AJ: **Genomics. Annotation of the *Drosophila* genome.** *Nature* 2001, **411:**259-260.
15. Gopal S, Schroeder M, Pieper U, Sczyrba A, Aytekin-Kurban G, Bekiranov S, Fajardo JE, Eswar N, Sanchez R, Sali A, *et al.*: **Homology-based annotation yields 1,042 new candidate genes in the *Drosophila melanogaster* genome.** *Nat Genet* 2001, **27:**337-340.
16. Andrews J, Bouffard GG, Cheadle C, Lu J, Becker KG, Oliver B: **Gene discovery using computational and microarray analysis of transcription in the *Drosophila melanogaster* testis.** *Genome Res* 2000, **10:**2030-2043.
17. Celniker SE, Wheeler DA, Kronmiller B, Carlson JW, Halpern A, Patel S, Adams M, Champe M, Dugan SP, Frise E, *et al.*: **Finishing a whole-genome shotgun: Release 3 of the *Drosophila* euchromatic genome sequence.** *Genome Biol* 2002, **3**:research0079.1-0079.14.
18. Kaminker JS, Bergman CM, Kronmiller B, Carlson J, Svirskas R, Patel S, Frise E, Wheeler DA, Lewis SE, Rubin GM, *et al.*: **The transposable elements of the *Drosophila melanogaster* euchromatin - a genomics perspective.** *Genome Biol* 2002, **3**:research0084.1-0084.20.
19. Lewis SE, Searle SMJ, Harris N, Gibson M, Iyer V, Richter J, Wiel C, Bayraktaroglu L, Birney E, Crosby L, *et al.*: **Apollo: a sequence annotation editor.** *Genome Biol* 2002, **3**:research0082.1-0082.14.
20. Reese MG, Kulp D, Tammana H, Haussler D: **Genie - gene finding in *Drosophila melanogaster*.** *Genome Res* 2000, **10:**529-538.
21. Burge C, Karlin S: **Prediction of complete gene structures in human genomic DNA.** *J Mol Biol* 1997, **268:**78-94.
22. **Wu-BLAST** [http://blast.wustl.edu]
23. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403-410.

24. Florea L, Hartzell G, Zhang Z, Rubin GM, Miller W: **A computer program for aligning a cDNA sequence with a genomic DNA sequence.** *Genome Res* 1998, **8:**967-974.
25. Hoskins RA, Smith CD, Carlson JW, Carvalho AB, Halpern A, Kaminker JS, Kennedy C, Mungall CJ, Sullivan BA, Sutton GG, *et al.*: **Heterochromatic sequences in a *Drosophila* whole-genome shotgun assembly.** *Genome Biol* 2002, **3:**research0085.1-0085.16.
26. The *C. elegans* Sequencing Consortium: **Genome sequence of the nematode *C. elegans*: a platform for investigating biology.** *Science* 1998, **282:**2012-2018.
27. Lin X, Kaul S, Rounsley S, Shea TP, Benito MI, Town CD, Fujii CY, Mason T, Bowman CL, Barnstead M, *et al.*: **Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*.** *Nature* 1999, **402:**761-768.
28. **Ensembl human genome browser** [http://www.ensembl.org/Homo_sapiens]
29. Stapleton M, Liao G, Brokstein P, Hong L, Carninci P, Shiraki T, Hayashizaki Y, Champe M, Pacleb J, Wan K, *et al.*: **The *Drosophila* gene collection: identification of putative full-length cDNAs for 70% of *D. melanogaster* genes.** *Genome Res* 2002, **12:**1294-1300.
30. Stapleton M, Carlson J, Brokstein J, Yu C, Champe M, George R, Guarin H, Kronmiller B, Pacleb J, Park S, *et al.*: **A *Drosophila* full-length cDNA resource.** *Genome Biol* 2002, **3:**research0080.1-0080-8.
31. Wilkin MB, Becker MN, Mulvey D, Phan I, Chao A, Cooper K, Chung HJ, Campbell ID, MacIntyre M, Macintyre R: ***Drosophila* dumpy is a gigantic extracellular protein required to maintain tension at epidermal-cuticle attachment sites.** *Curr Biol* 2000, **10:**559-567.
32. Lowe TM, Eddy SR: **tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence.** *Nucleic Acids Res* 1997, **25:**955-964.
33. Wooley JC, Cone RD, Tartof D, Chung SY: **Small nuclear ribonucleoprotein complexes of *Drosophila melanogaster*.** *Proc Natl Acad Sci USA* 1982, **79:**6762-6766.
34. Tycowski KT, Steitz JA: **Non-coding snoRNA host genes in *Drosophila*: expression strategies for modification guide snoRNAs.** *Eur J Cell Biol* 2001, **80:**119-125.
35. Lowe TM, Eddy SR: **A computational screen for methylation guide snoRNAs in yeast.** *Science* 1999, **283:**1168-1171.
36. Shendure J, Church GM: **Computational discovery of sense-antisense transcription in the human and mouse genomes.** *Genome Biol* 2002, **3:**research0044.1-0044.14.
37. Echols N, Harrison P, Balasubramanian S, Luscombe NM, Bertone P, Zhang Z, Gerstein M: **Comprehensive analysis of amino acid and nucleotide composition in eukaryotic genomes, comparing genes and pseudogenes.** *Nucleic Acids Res* 2002, **30:**2515-2523.
38. Mounsey A, Bauer P, Hope IA: **Evidence suggesting that a fifth of annotated *Caenorhabditis elegans* genes may be pseudogenes.** *Genome Res* 2002, **12:**770-775.
39. Stein L, Sternberg P, Durbin R, Thierry-Mieg J, Spieth J: **WormBase: network access to the genome and biology of *Caenorhabditis elegans*.** *Nucleic Acids Res* 2001, **29:**82-86.
40. **WormBase, The Genome and Biology of *C. elegans*, Release WS88** [http://www.wormbase.org]
41. Gu Z, Cavalcanti A, Chen FC, Bouman P, Li WH: **Extent of gene duplication in the genomes of *Drosophila*, nematode, and yeast.** *Mol Biol Evol* 2002, **19:**256-262.
42. De Gregorio E, Spellman PT, Rubin GM, Lemaitre B: **Genome-wide analysis of the *Drosophila* immune response by using oligonucleotide microarrays.** *Proc Natl Acad Sci USA* 2001, **98:**12590-12595.
43. Krauss V, Reuter G: **Two genes become one: the genes encoding heterochromatin protein *Su(var)3-9* and translation initiation factor subunit eIF-2gamma are joined to a dicistronic unit in holometabolic insects.** *Genetics* 2000, **156:**1157-1167.
44. Ashburner M, Misra S, Roote J, Lewis SE, Blazej R, Davis T, Doyle C, Galle R, George R, Harris N, *et al.*: **An exploration of the sequence of a 2.9-Mb region of the genome of *Drosophila melanogaster*: the *Adh* region.** *Genetics* 1999, **153:**179-219.
45. Spencer CA, Gietz RD, Hodgetts RB: **Overlapping transcription units in the *dopa decarboxylase* region of *Drosophila*.** *Nature* 1986, **322:**279-281.
46. Kumar M, Carmichael GG: **Antisense RNA: function and fate of duplex RNA in cells of higher eukaryotes.** *Microbiol Mol Biol Rev* 1998, **62:**1415-1434.
47. Vanhee-Brossollet C, Vaquero C: **Do natural antisense transcripts make sense in eukaryotes?** *Gene* 1998, **211:**1-9.
48. Labrador M, Mongelard F, Plata-Rengifo P, Baxter EM, Corces VG, Gerasimova TI: **Protein encoding by both DNA strands.** *Nature* 2001, **409:**1000.
49. Dorn R, Reuter G, Loewendorf A: **Transgene analysis proves mRNA trans-splicing at the complex *mod(mdg4)* locus in *Drosophila*.** *Proc Natl Acad Sci USA* 2001, **98:**9724-9729.
50. Giniger E, Tietje K, Jan LY, Jan YN: ***lola* encodes a putative transcription factor required for axon growth and guidance in *Drosophila*.** *Development* 1994, **120:**1385-1398.
51. Read D, Butte MJ, Dernburg AF, Frasch M, Kornberg TB: **Functional studies of the BTB domain in the *Drosophila* GAGA and Mod(mdg4) proteins.** *Nucleic Acids Res* 2000, **28:**3864-3870.
52. Bell AC, West AG, Felsenfeld G: **Insulators and boundaries: versatile regulatory elements in the eukaryotic genome.** *Science* 2001, **291:**447-450.
53. Harvey AJ, Bidwai AP, Miller LK: **Doom, a product of the *Drosophila mod(mdg4)* gene, induces apoptosis and binds to baculovirus inhibitor-of-apoptosis proteins.** *Mol Cell Biol* 1997, **17:**2835-2843.
54. Pauli D, Tonka CH, Ayme-Southgate A: **An unusual split *Drosophila* heat shock gene expressed during embryogenesis, pupation and in testis.** *J Mol Biol* 1988, **200:**47-53.
55. Schulz RA, Miksch JL, Xie XL, Cornish JA, Galewsky S: **Expression of the *Drosophila gonadal* gene: alternative promoters control the germ-line expression of monocistronic and bicistronic gene transcripts.** *Development* 1990, **108:**613-622.
56. Andrews J, Smith M, Merakovsky J, Coulson M, Hannan F, Kelly LE: **The *stoned* locus of *Drosophila melanogaster* produces a dicistronic transcript and encodes two distinct polypeptides.** *Genetics* 1996, **143:**1699-1711.
57. Brogna S, Ashburner M: **The *Adh-related* gene of *Drosophila melanogaster* is expressed as a functional dicistronic messenger RNA: multigenic transcription in higher organisms.** *EMBO J* 1997, **16:**2023-2031.
58. Niimi T, Yokoyama H, Goto A, Beck K, Kitagawa Y: **A *Drosophila* gene encoding multiple splice variants of Kazal-type serine protease inhibitor-like proteins with potential destinations of mitochondria, cytosol and the secretory pathway.** *Eur J Biochem* 1999, **266:**282-292.
59. Gray TA, Nicholls RD: **Diverse splicing mechanisms fuse the evolutionarily conserved bicistronic MOCS1A and MOCS1B open reading frames.** *RNA* 2000, **6:**928-936.
60. Liu H, Jang JK, Graham J, Nycz K, McKim KS: **Two genes required for meiotic recombination in *Drosophila* are expressed from a dicistronic message.** *Genetics* 2000, **154:**1735-1746.
61. Walker DL, Wang D, Jin Y, Rath U, Wang Y, Johansen J, Johansen KM: **Skeletor, a novel chromosomal protein that redistributes during mitosis provides evidence for the formation of a spindle matrix.** *J Cell Biol* 2000, **151:**1401-1412.
62. Levine F, Yee JK, Friedmann T: **Efficient gene expression in mammalian cells from a dicistronic transcriptional unit in an improved retroviral vector.** *Gene* 1991, **108:**167-174.
63. Peabody DS, Berg P: **Termination-reinitiation occurs in the translation of mammalian cell mRNAs.** *Mol Cell Biol* 1986, **6:**2695-2703.
64. Oh SK, Scott MP, Sarnow P: **Homeotic gene *Antennapedia* mRNA contains 5′-noncoding sequences that confer translational initiation by internal ribosome binding.** *Genes Dev* 1992, **6:**1643-1653.
65. Bergman CM, Pfeiffer BD, Rincón-Limas DE, Hoskins RA, Gnirke A, Mungall CJ, Wang AM, Kronmiller B, Pacleb J, Park S, *et al.*: **Assessing the impact of comparative genomic sequence data on the functional annotation of the *Drosophila* genome.** *Genome Biol* 2002, **3:**research0086.1-0086.20.
66. Reese MG, Hartzell G, Harris NL, Ohler U, Abril JF, Lewis SE: **Genome annotation assessment in *Drosophila melanogaster*.** *Genome Res* 2000, **10:**483-501.
67. Haas BJ, Volfovsky N, Town CD, Troukhan M, Alexandrov N, Feldmann KA, Flavell RB, White O, Salzberg SL: **Full-length messenger RNA sequences greatly improve genome annotation.** *Genome Biol* 2002, **3:**research0029.1-0029.12.

68. Gaasterland T, Oprea M: **Whole-genome analysis: annotations and updates.** *Curr Opin Struct Biol* 2001, **11:**377-381.
69. **Expressed Sequence Tags Database (dbEST)** [http://www.ncbi.nlm.nih.gov/dbEST]
70. Kans JA, Ouellette BF: **Submitting DNA sequences to the databases.** *Methods Biochem Anal* 2001, **43:**65-81.
71. **Sequin - a DNA sequence submission and update tool** [http://www.ncbi.nlm.nih.gov/Sequin]
72. Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, Bucher P, Cerutti L, Corpet F, Croning MD, *et al.*: **InterPro - an integrated documentation resource for protein families, domains and functional sites.** *Bioinformatics* 2000, **16:**1145-1150.
73. Spradling AC, Stern D, Beaton A, Rhem EJ, Laverty T, Mozden N, Misra S, Rubin GM: **The Berkeley *Drosophila* Genome Project gene disruption project: single P-element insertions mutating 25% of vital *Drosophila* genes.** *Genetics* 1999, **153:**135-177.
74. **Apollo gene annotation tool: developer page** [http://www.fruitfly.org/annot/apollo]
75. **BDGP: download BDGP sequence and annotation databases** [http://www.fruitfly.org/sequence/download.html]
76. **FlyBase update form** [http://www.fruitfly.org/cgi-bin/annot/FBupdate]
77. Ohler U, Liao G-C, Niemann H, Rubin GM: **Computational analysis of core promoters in the *Drosophila* genome.** *Genome Biol* 2002, **3:**research0087.1-0087.12 .
78. **Gene Ontology Consortium** [http://www.geneontology.org]
79. Mi H, Vandergriff J, Campbell MJ, Marechania A, Majoros W, Lewis S, Thomas PD, Ashburner M: **Assessment of genome-wide protein function classification for *Drosophila melanogaster.*** *Genome Res*, in press.
80. **SWISS-PROT** [http://www.ebi.ac.uk/swissprot]
81. **TrEMBL** [http://www.ebi.ac.uk/trembl/index.html]
82. **BDGP re-annotation guidelines** [http://www.fruitfly.org/annot/reannot-guidelines.html]
83. **SP-TrEMBL (SWALL, SWISS-PROT/TrEMBL/TrEMBLNEW) database** [http://www.ebi.ac.uk/trembl]
84. **UniGene *Mus musculus*** [http://www.ncbi.nlm.nih.gov/UniGene/query.cgi?ORG=Mm]
85. **TrEMBL Protein Database Statistics** [http://www2.ebi.ac.uk/swissprot/sptr_stats]
86. Zdobnov EM, Von Mering C, Letunic I, Torrents D, Suyama M, Copley RR, Christophides GK, Thomasova D, Holt RA, Subramanian GM, *et al.*: **Comparative genome and proteome analysis of *Anopheles gambiae* and *Drosophila melanogaster.*** *Science* 2002, **298:**149-159.
87. **GFF format: an exchange format for feature description** [http://www.sanger.ac.uk/Software/formats/GFF]